A Cognitive and Computational Model of the Sound Change from Pre- to Post-Aspiration in Andalusian Spanish

Johanna Cronenberg

Munich, 2024

A Cognitive and Computational Model of the Sound Change from Pre- to Post-Aspiration in Andalusian Spanish

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
der Ludwig-Maximilians-Universität München

vorgelegt von
Johanna Cronenberg
aus Bonn

2024

Erstgutachter: Prof. Jonathan Harrington
Zweitgutachter: Prof. James Kirby
Drittgutachter: PD Dr. Christoph Draxler
Tag der mündlichen Prüfung: 14.02.2023

Für Martha & Peter Cronenberg
Et hätt noch emmer joot jejange.

Zusammenfassung

Eines der auffälligsten phonetischen Merkmale, das andalusisches vom kastilischen (Standard) Spanisch unterscheidet, ist die Glottalisierung bzw. Abschwächung von /s/ zu /h/ (Canfield, 1981; Torreira, 2006; Villena-Ponsoda, 2008). In Clustern von /s/ gefolgt von einem stimmlosen Plosiv ist der Plosiv folglich prä-aspiriert, z.B. *resto* /resto/ > /re^{h}to/ (dt. Rest). In der jüngeren Vergangenheit haben jedoch mehrere Studien gezeigt, dass ein Lautwandel hin zur Post-Aspiration stattfindet, z.B. /re^{h}to/ > /ret^{h}o/ (Parrell, 2012; Ruch & Harrington, 2014; Torreira, 2012). Als synchrone Ursache dieses Lautwandels wird eine Resynchronisierung artikulatorischer Gesten bei schnellem Sprechtempo vermutet (Parrell, 2012). Die Koordination der Glottisöffnung mit dem oralen Verschluss bestimmt, ob der stimmlose Plosiv unaspiriert (beide Gesten müssen synchronisiert und gleich lang sein), prä-aspiriert (der Verschluss wird nach der Glottisöffnung geformt), oder post-aspiriert ist (der Verschluss wird gelöst, während die Glottis noch geöffnet ist). Diese Dissertation beschäftigt sich mit dem Lautwandel von Prä- zu Post-Aspiration im andalusischen Spanisch, aber hofft auch einen Beitrag zu unserem allgemeinen Verständnis von Lautwandeln leisten zu können. Hierzu werden neue Ansätze präsentiert, die sowohl den Ursprung von Lautwandeln in der Dynamik gesprochener Sprache als auch deren Verbreitung durch interindividuelle Interaktionen in Betracht ziehen. Das erste Ziel dieser Arbeit war es, den Lautwandel in einer Art und Weise zu analysieren, die die zugrundeliegende Resynchronisierung der Gesten widerspiegelt. Die dynamische Analysemethode, die zu diesem Zweck entwickelt wurde, ermöglicht die Beobachtung phonetischer Details, die eine wichtige Rolle in Lautwandeln spielen, aber in traditionellen Analysen verborgen bleiben können. Da das Aufkommen und die Verbreitung von Lautwandeln von einem komplexen Zusammenspiel von Faktoren abhängt, können computationelle Simulationen ein hilfreiches Werkzeug sein, um Erkenntnisse zu gewinnen. Deshalb war das zweite Ziel dieser Dissertation die Entwicklung eines agent-basierten Modells (ABM), das zeigt, wie Lautwandel aus der Verstärkung von phonetischen Biases und der Reorganisation phonologischer

Klassen entstehen kann. Dieses ABM wurde dann eingesetzt, um den Wandel von Prä- zu Post-Aspiration im andalusischen Spanisch zu modellieren.

Die vorliegende Arbeit besteht aus drei Hauptkapiteln, die im Folgenden kurz zusammengefasst werden. In Kapitel 2 (publiziert als Cronenberg et al., 2020) wurde der Lautwandel von Prä- zu Post-Aspiration mittels einer neuen Methode untersucht, die auf akustischen Signalen anstelle einer vertikalen Segmentierung des Sprachsignals beruht (Fowler, 1984). Die akustischen Signale spiegelten die glottale und orale Geste wider: die orale Geste wurde durch die Stimmhaftigkeitswahrscheinlichkeit repräsentiert (VP; Gonzalez & Brookes, 2014), während der Verschlussgrad und das Aspirationsrauschen vom hoch-frequenten Energiesignal (HF) abgelesen wurden. Diese Signale wurden für die intervokalischen Cluster /sp, st, sk/ in Sprachaufnahmen von 48 andalusischen Sprechern berechnet und anschließend der Functional Principal Components Analysis (FPCA) übergeben. FPCA ist eine Analysetechnik, die die Variationen in den Signalformen ausgibt (Principal Components, PCs). Der erste PC fing die Synchronisierung des Verschlusses im Verhältnis zum stimmlosen Intervall ein, was zeigte, dass Prä- und Post-Aspiration durch das Timing des Verschlusses in einer wechselseitigen Beziehung standen. Zudem wurde zum ersten Mal festgestellt, dass der sogenannte Trade-Off zwischen Prä- und Post-Aspiration auch mittels der Amplitude des Aspirationsrauschens signalisiert wurde. Der erste PC unterschied auch zwischen den Sprechergruppen: jüngere SprecherInnen produzierten /sp, st, sk/ mit mehr Post-Aspiration als ältere SprecherInnen, die wiederum mehr Prä-Aspiration benutzten. Der Lautwandel war außerdem weiter fortgeschritten in West- als in Ost-Andalusien. Anhand des Apparent-Time Paradigmas (Bailey et al., 1991) etablierten diese Ergebnisse eine Verbindung zwischen der synchronen Basis (d.h. der Resynchronisierung artikulatorischer Gesten) und einem fortschreitenden Lautwandel. Im Hinblick auf die menschliche Sprachverarbeitung und Lautwandel im Allgemeinen deutete die Studie in Kapitel 2 auf zwei Arten von Informationen hin, die aus dem Sprachsignal extrahiert werden können. Die erste Art is phonologisches Wissen, das etwas über die Charakteristika von Lautkategorien im Vergleich zu anderen innerhalb einer Sprachgemeinschaft aussagt. Zum Beispiel werden /sp, st, sk/ aspiriert (wo genau die Aspira-

tion produziert wird, spielt keine Rolle), während /p, t, k/ unaspiriert sind. Phonologisches Wissen wird abstrahiert von gespeicherten Sprachsignalen (Pierrehumbert, 2001) und enthält auch Informationen über die artikulatorischen Gesten und ihre Synchronisierung. Die zweite Art von Informationen betrifft die Verteilung bestimmter Aussprachevarianten, d.h. die Eigenschaften einer SprecherIn wie z.B. das Alter oder die Herkunft korrelieren mit der Produktion bestimmter Laute. Aus diesem Grund besteht bei HörerInnen, die eine Vielzahl von Kontakten zu SprecherInnen haben, eine höhere Wahrscheinlichkeit, dass sie einen perzeptiven Trade-Off bspw. zwischen Prä- und Post-Aspiration entwickeln (Beddor, 2009). Die Schlussfolgerung aus Kapitel 2 ist, dass eine Transformation (wie z.B. FPCA) wahrgenommener und sozial variabler Sprachsignale in der menschlichen Sprachverarbeitung angewendet wird, um Informationen über Phonologie und Verteilung daraus abzuleiten.

Kapitel 3 behandelt das auf kognitiven Prinzipien beruhende agent-basierte Lautwandelmodell, das als Teil dieser Dissertation entwickelt wurde. Obwohl Lautwandel seit Jahrhunderten erforscht wird, ist noch Vieles über das Zusammenspiel intra- und extra-linguistischer Faktoren, die zum Entstehen und zur Verbreitung von Lautwandeln beitragen, unbekannt. Computationelle agent-basierte Modelle (ABMs) bieten ein kontrolliertes Umfeld, in dem die Rolle kognitiver, sozialer und linguistischer Faktoren in Lautwandeln untersucht werden kann. In ABMs interagieren Agenten miteinander, indem sie einer Reihe an Regeln folgen, und es kann in der Folge beobachtet werden, wie individuelle Entscheidungen (*micromotives*) zu globalen Mustern oder Verhaltensweisen führen (*macrobehaviours*; Schelling, 1978). Das ABM namens `soundChangeR`, das hier präsentiert wird, ist das erste seiner Art, das als frei verfügbares R Paket implementiert und vollständig dokumentiert wurde (https://github.com/IPS-LMU/soundChangeR). Es basiert auf dem theoretischen interaktiven-phonetischen (IP) Modell (Harrington et al., 2018), welches davon ausgeht, dass Lautwandel das Ergebnis stochastischer Interaktionen zwischen Individuen ist, durch die bereits existierende phonetische Biases verstärkt werden. Das IP Modell übernimmt von der Exemplar Theorie und episodischen Modellen des Gedächtnisses (Goldinger, 1996; Goldinger & Azuma, 2004; Pierrehumbert, 2001, 2003a), dass Sprachexemplare mitsamt phone-

tischer Details abgespeichert werden und phonologische Klassen sich von diesen Exemplarwolken ableiten lassen. Wenn das Individuum mehr Spracherfahrungen gesammelt und dementsprechend mehr Exemplare abgespeichert hat, werden die phonologischen Klassen ggf. umstrukturiert (Norris et al., 2003; Pardo et al., 2012). Jeder Agent wird mit parametrisierten Sprachexemplaren initialisiert, die von einer menschlichen SprecherIn produziert wurden, bevor die Interaktionen beginnen. In einer Interaktion extrahiert der sprechende Agent ein neues Token eines Wortes aus einer Normalverteilung, die aus den abgespeicherten Exemplaren gebildet wird. Das neue Token wird zusammen mit dem assoziierten Worttyp dem hörenden Agenten übermittelt, der entscheidet, ob er das perzipierte Token abspeichert. Diese Entscheidung kann entweder anhand der Typikalität (d.h. ob das Token typisch genug für das intendierte Phonem ist) oder Unterscheidbarkeit (d.h. ob das Token wahrscheinlicher zum intendierten als zu konkurrierenden Phonemen gehört) des Tokens getroffen werden (Todd et al., 2019). Es wurde ein künstlicher Datensatz erstellt, um den Einfluss der Speicherkriterien auf das Simulationsergebnis zu untersuchen. Der Datensatz bestand aus zwei Agenten, A und B, die eine ähnliche Phonemklasse SP2, aber unterschiedliche Phonemklassen SP1 hatten, wobei SP1 von Agent A breiter und zu SP1 von Agent B ausgerichtet war. In einer Simulation, in der nur das Typikalitätskriterium angewendet wurde, verschob sich SP1 von Agent A im akustischen Raum hin zu dem von B, während Agent Bs Phonemklassen sich nicht veränderten und eine leichte Überlappung beibehielten. Als jedoch nur das Unterscheidbarkeitskriterium angewendet wurde, adaptierte Agent B die Phonemklasse SP1 von Agent A, sodass beide Agenten am Ende zwei nicht überlappende Phoneme SP1 aufwiesen. SP2 veränderte sich bei keinem Agenten und in keiner der beiden Simulationen. Diese Simulationen zeigten, dass `soundChangeR` sowohl phonetische Veränderungen als auch Stabilität simulieren kann (Sóskuthy, 2015). Da Phoneme eine wichtige Rolle in der Perzeption der Agenten spielen, bietet `soundChangeR` Algorithmen an, die Agent-spezifische Sub-Phonemklassen von den gespeicherten Exemplaren und deren Assoziation zu Wortklassen ableiten können. Mithilfe eines weiteren künstlichen Datensatzes, der nur aus einem Agenten bestand, wurde gezeigt, wie diese Algorithmen Sub-Phoneme

identifizieren. Bei einer systematischen Beziehung zwischen gespeicherten Exemplaren und Wortklassen (d.h. Exemplare der Wörter W1 bis W5 waren in einem anderen Teil des akustischen Raums angesiedelt als Exemplare der Wörter W6 bis W10) wurden zwei Sub-Phoneme identifiziert. Bei einer zufälligen Beziehung zwischen denselben Exemplaren und Wortklassen bestimmte das Phonologiemodul in `soundChangeR` hingegen nur ein Sub-Phonem. Zusätzlich fanden Gubian et al. (2023), dass dieses flexible Phonologiemodul adäquat den Zusammenfall des phonologischen Kontrastes zwischen /ɪə/ und /eə/ in Neuseeland-Englisch modellierte, während es die phonologische Stabilität bewahrte, die zur Modellierung von /u/-fronting in Standard Southern British English vonnöten war. Im letzten Teil des Kapitels 3 wird das ABM `soundChangeR` mit anderen computationellen Lautwandelmodellen verglichen und es werden verschiedene Erweiterungen der Software diskutiert.

Kapitel 4 kombiniert die dynamische Analyse aus Kapitel 2 mit dem agent-basierten Modell aus Kapitel 3, um den Lautwandel von Prä- zu Post-Aspiration in andalusischen Clustern zu simulieren. Im ersten Teil dieser Studie wird ein Datensatz erstellt, der Wörter mit /st/ und /t/ enthählt, wobei /t/ als phonologisch unaspiriert gilt. Dementsprechend bildet /t/ einen phonologischen Kontrast zu /st/, der trotz des Lautwandels hin zur Post-Aspiration in /st/ bestehen bleiben sollte, und dient somit als Plausibilitätsprüfung für die Simulation. Dieselben 48 SprecherInnen wie in Kapitel 2 produzierten 19 Wörter mit /st/ und 9 mit /t/, für die anschließend die Stimmhaftigkeitswahrscheinlichkeit und das Energiesignal für hohe Frequenzen abgeleitet wurden. Auf diese Signalpaare wurde FPCA angewendet und die ersten vier PCs wurden analysiert. PC1 und vor allem PC4 unterschieden /st/ und /t/ voneinander und zeigten, dass das Cluster durch ein längeres stimmloses Intervall, mehr hochfrequente Energie und daher auch mehr Aspiration gekennzeichnet war. Auch /t/ wurde mit ein wenig Post-Aspiration produziert, was zwar nicht zu seiner phonologischen Beschreibung passt, aber auch nicht ganz unerwartet ist, da die plötzliche Lösung des oralen Verschlusses für einige Millisekunden die angestaute Luft entweichen lässt, bevor die Stimmlippen wieder anfangen zu schwingen. PC2 beschrieb die Position der Aspirationsphase in /st/, d.h. ob das Cluster prä- oder post-aspiriert war (oder eine Mischung von beidem).

Diese Information korrelierte auch mit dem Alter der SprecherInnen: jüngere SprecherInnen tendierten zu mehr Post-Aspiration, während ältere SprecherInnen die Cluster eher prä-aspirierten. Dieses Ergebnis unterstützt frühere Studien, die einen Wandel hin zu Post-Aspiration fanden. Mit dem Ziel, diesen Lautwandel zu modellieren, wurden die PC Scores aus diesem ersten Teil der Studie als Input für eine Simulation mit `soundChangeR` im zweiten Teil verwendet. Die Agenten in dieser Simulation repräsentierten ältere und jüngere andalusische SprecherInnen, die Exemplare von /st/ und /t/ austauschten. Diese beiden kanonischen Phoneme wurden wie erwartet korrekt identifiziert und für die Dauer der Simulation vom Phonologiemodul in `soundChangeR` beibehalten. Auf dem akustischen Level produzierten alle Agenten /t/ nach den Interaktionen genauso wie zuvor, aber die jüngeren und älteren Agenten konvergierten entgegen der Erwartungen zu einer Variante von /st/, die sowohl Prä- als auch Post-Aspiration enthielt. Dieses Ergebnis ist wahrscheinlich zustande gekommen, weil die PC Scores der älteren Agenten keine Verzerrung hin zu denen der jüngeren Agenten zeigten. Solche Verzerrungen in Kombination mit der selektiven Speicherung von Exemplaren sind jedoch einer der Wirkungsmechanismen in `soundChangeR`, die solche akustischen Veränderungen auslösen können (s. 3.3.2). Außerdem wurden im letzten Teil von Kapitel 4 zwei weitere Gründe für die unrealistischen Ergebnisse der Simulation diskutiert. Der eine ist die generelle Tendenz von `soundChangeR`, die sich gegen akustisch extreme Exemplare richtet. Dies könnte ein Problem darstellen, da solche Outlier besonders auffällig sind und daher eine wichtige Rolle in Lautwandeln spielen könnten. Der andere Grund ist, dass `soundChangeR` keine sogenannten perzeptiven Cues modellieren kann, deren Gewichtung sich vor allem in Lautwandeln, bei denen neue Laute phonologisiert werden, ändern kann.

Abstract

One of the most salient characteristics that distinguishes Andalusian Spanish from Castilian (Standard) Spanish phonetically is that /s/ is glottalised or lenited to /h/ (Canfield, 1981; Torreira, 2006; Villena-Ponsoda, 2008). In clusters of /s/ followed by a voiceless plosive, the plosive is thus pre-aspirated, e.g. *resto* /resto/ > /re^{h}to/ (engl. rest). In recent years, several studies have shown that there is an ongoing change from pre- to post-aspiration, e.g. /re^{h}to/ > /ret^{h}o/ (Parrell, 2012; Ruch & Harrington, 2014; Torreira, 2012). The synchronic basis of this sound change is believed to be a resynchronisation of articulatory gestures at faster speech rates (Parrell, 2012). More specifically, the way that the glottal opening is aligned with the oral closure determines whether a voiceless plosive is unaspirated (both gestures must be synchronised and of the same duration), pre-aspirated (the closure is formed after the glottis has been opened), or post-aspirated (the closure is released while the glottis is still open). This thesis is concerned with the sound change by which pre- becomes post-aspiration in Andalusian Spanish, but also hopes to contribute more generally to our understanding of sound changes. This was attempted by presenting new approaches that take into account both the origins of sound change in the dynamics of spoken language as well as its spread through interactions between individuals. More specifically, the first aim of this thesis was to analyse the change in a way that reflected the underlying gestural resynchronisation. The dynamic analysis developed for that purpose allows for the observation of phonetic details that play an important role in sound changes but might be easily missed in traditional analyses. Since the emergence and spread of sound changes depend on a complex interplay of factors, computational simulations can be a helpful tool to gain insights. Therefore, the second aim of this thesis was to develop an agent-based model (ABM) which shows how sound changes can emerge from the reinforcement of phonetic biases and reorganisation of phonological classes. This ABM was then used to model the change from pre- to post-aspiration in Andalusian Spanish.

This thesis consists of three main chapters which are briefly summarised here. In chapter 2 (published as Cronenberg et al., 2020), the change from pre- to post-aspiration was investigated by means of a novel method that uses acoustic signals as proxies for the glottal and oral gestures instead of relying on a vertical segmentation of the speech signal (Fowler, 1984). The glottal gesture was represented by the voicing probability (VP; Gonzalez & Brookes, 2014); the degree of closure as well as aspiration noise was represented by the high-frequency energy signal (HF). These signals, computed for intervocalic clusters /sp, st, sk/ in the speech of 48 Andalusian Spanish speakers, were supplied to Functional Principal Components Analysis (FPCA), a technique that returns the main modes of variation in the signals' shapes (Principal Components, PCs). The first PC captured the phasing of the closure in relation to the voiceless interval which showed that pre- and post-aspiration were inversely related to each other through the timing of the closure. Moreover, it was shown for the first time that the trade-off between pre- and post-aspiration was also signalled by the amplitude of the aspiration noise. The first PC also distinguished between the speaker groups: younger speakers were shown to produce /sp, st, sk/ with more post-aspiration than older speakers who predominantly used pre-aspiration; the sound change was also further advanced in speakers from West than East Andalusia. These results established a link between the synchronic basis (i.e. the resynchronisation of articulatory gestures) and a sound change in progress as shown by the apparent-time approach (Bailey et al., 1991). With regard to both human speech processing and sound change, the study in chapter 2 suggested that two kinds of information can be extracted from the dynamic speech signal. The first is phonological knowledge which indicates population-level characteristics of sound categories as compared to others, e.g. that /sp, st, sk/ are produced with aspiration surrounding the closure whereas /p, t, k/ are not. Phonological knowledge, learned and abstracted from memorised traces of speech, also comprises information about articulatory gestures and their alignment. The second kind of information is distributional, i.e. characteristics of a speaker such as their age or regional origin correlate with the way they produce certain speech sounds. Thus, listeners who have been exposed to a wide variety of speakers are more likely to develop

a perceptual trading relationship between e.g. pre- and post-aspiration. While it has been suggested previously that phonological knowledge is a statistical abstraction over memorised exemplars (Pierrehumbert, 2001) and that trading relationships can be a precursor to sound change (Beddor, 2009), the new angle in chapter 2 is that a transformation (such as FPCA) of the perceived time-varying and socially variable speech material is used to derive knowledge from it that otherwise cannot be extracted.

Chapter 3 is about the cognitively-inspired agent-based model of sound change which was developed as part of this thesis. Even though sound change has inspired research for centuries, the complex interplay between intra- and extralinguistic factors that may contribute to its initiation and propagation remains poorly understood. Computational agent-based models (ABMs) offer a controlled environment in which the role of cognitive, social, and linguistic factors in sound change can be explored. In ABMs, agents interact with one another along a defined set of rules and it can be observed how individual actions (micromotives) lead to population-wide patterns (macrobehaviours; Schelling, 1978). The ABM presented here, called `soundChangeR`, is the first of its kind that was implemented as a publicly available R package and comes with a full documentation of the code (https://github.com/IPS-LMU/soundChangeR). It is based on the theoretical interactive-phonetic (IP) model (Harrington et al., 2018) which assumes that sound changes are the result of stochastic interactions between individuals through which existing phonetic biases are magnified. The IP model takes from exemplar theory and episodic models of memory (Goldinger, 1996; Goldinger & Azuma, 2004; Pierrehumbert, 2001, 2003a) that traces of speech (exemplars) are stored and phonological classes are abstractions over clouds of exemplars which can occasionally be regrouped when the individual has acquired more language experience (Norris et al., 2003; Pardo et al., 2012). Each agent in `soundChangeR` is initialised with parameterised traces of speech (exemplars) produced by an actual speaker before starting to interact with other agents. In an interaction, an agent speaker uses a Gaussian sampling procedure to produce a new token of a word which is transmitted to an agent listener whose task it is to decide whether or not to memorise the token. This decision can be based either on the token's typicality

(i.e. is it a typical enough member of the intended phoneme?) or discriminability (i.e. is it more likely to belong to the intended phoneme than to all competing phonemes?; also see Todd et al., 2019). An artificial dataset was used to demonstrate the impact of the memorisation criteria on the simulation's outcome. The dataset consisted of two agents, A and B, who had a similar representation of a phoneme SP2, but different representations of a phoneme SP1 such that agent A's SP1 was broader and skewed towards that of agent B. In a simulation in which only the typicality criterion was applied, agent A's SP1 shifted towards that of agent B in the acoustic space, whereas agent B's phoneme classes did not change and maintained a slight overlap. When only the discriminability criterion was applied, on the other hand, agent B's SP1 became more like that of agent A, leaving both agents with two non-overlapping phoneme classes. SP2 did not change in either simulation. These simulations showed that `soundChangeR` is capable of modelling both phonetic changes and stability (Sóskuthy, 2015). Since phonemes play an important role in the agent's perception, `soundChangeR` provides algorithms that can derive agent-specific sub-phonemic classes from the stored exemplars and their fixed association to word labels. Using another artificial dataset consisting of only one agent, it was shown how these algorithms identify sub-phonemic classes. When the association between stored exemplars and word labels was systematic (i.e. exemplars of words W1 to W5 were in a different part of the acoustic space than those of words W6 to W10), two sub-phonemic classes were identified. When the exemplars' association to word classes was random, on the other hand, the flexible phonology module in `soundChangeR` determined that there was only one sub-phoneme. Additionally, Gubian et al. (2023) found that this flexible phonology module can adequately model phonological change in the New Zealand English merger of /ɪə/ and /eə/ while maintaining the phonological stability necessary to model /u/-fronting in Standard Southern British English. In the final part of chapter 3, the ABM `soundChangeR` is compared to other computational models of sound change and several possible extensions of the software are considered and discussed.

Chapter 4 combines the dynamic analysis technique from chapter 2 and the agent-based model presented in chapter 3 to simulate the change from

pre- towards post-aspiration in Andalusian Spanish clusters. In the first part, a dataset is composed of words containing either /st/ or /t/, the latter of which is considered to be phonologically unaspirated. Thus, /t/ posits a phonological contrast to /st/ which should not be affected by the sound change towards post-aspiration and can therefore serve as a sanity check in the simulation. The same 48 Andalusian Spanish speakers that provided the speech material for chapter 2 produced 19 words containing /st/ and 9 words containing /t/ from which the high-frequency energy and voicing probability signals were extracted between the two vowels on either side of the voiceless plosive. FPCA was applied on these signal pairs and the first four Principal Components were analysed. PC1 and, more clearly, PC4 separated /st/ from /t/ and showed that the cluster was characterised by a longer voiceless interval, more high-frequency energy, and therefore more overall aspiration. Nevertheless, /t/ was produced with a small amount of post-aspiration which does not match its phonological description, but is also not unexpected given that the release of the oral closure lets air escape with some force for a few milliseconds before the vocal folds can start swinging again. PC2 captured the location of the aspiration phase in /st/, i.e. whether the cluster was pre- or post-aspirated (or a mixture of both). This information also correlated significantly with the speakers' age: younger speakers tended to post-aspirate while older speakers were more likely to pre-aspirate the cluster. This result supported earlier studies that found a sound change in progress towards post-aspiration. With the aim of modelling the change by which pre- gives way to post-aspiration, the PC scores extracted in the first part of chapter 4 were used as input to a simulation with `soundChangeR` in the second part. The agents in this simulation represented older and younger Andalusian Spanish speakers who exchanged exemplars of /st/ and /t/. These two canonical phonemes were correctly identified and maintained by the flexible phonology module in `soundChangeR`, as expected. On the acoustic level, all agents produced /t/ the same way after as compared to before the interactions, but younger and older agents converged towards a common variant of /st/ that was characterised by both pre- and post-aspiration. This result can be attributed to the lack of skew in the older agents' PC scores towards those of the younger agents. Skew and orientation

of phonemic classes in combination with selective memorisation are a key mechanism in soundChangeR which can trigger acoustic changes, as shown in section 3.3.2. Two further possible reasons for the ABM's failure to accurately model the change from pre- to post-aspiration in Andalusian Spanish are discussed in the final part of chapter 4. The first is soundChangeR's general bias against extreme exemplars which might pose a problem given that such outliers are particularly salient and might play an important role in sound changes. The second is that soundChangeR cannot model the re-weighting of perceptual cues which is considered an important part of sound changes which involve phonologisation.

Acknowledgements

First and foremost I would like to thank my supervisor, Prof. Jonathan Harrington. He gave me the opportunity to do a PhD in phonetics – and it was all I had hoped it would be. Thank you, Jonathan, for all of the insightful discussions, your quick and most constructive feedback, and continuing support and encouragement. Under your supervision, I have written papers, given talks at conferences, taught classes, and gotten to know so many great people with whom I share the passion for phonetics. You have made me the researcher I am today.

Prof. James Kirby has provided me with the luxury of financial security so that I could finish this thesis. For being patient, generous, and for kindly inviting me into your research group: thank you, James. I am very much looking forward to working with you.

My closest co-worker during my PhD was, without a doubt, Michele Gubian. We have supported each other through the full range of emotions associated with software development and research (joy, desperation, excitement, rage). Michele, I'd be grateful if at least a bit of your curiosity and analytical mind have rubbed off on me over the years. Thank you for being such a great colleague and teacher.

It shows how dedicated of a teacher Florian Schiel is that he agreed to teach a class in which I was the only student. In that class, he introduced me to the agent-based model, which he had originally developed. Florian, you have sparked my interest in this field in a way I never would have imagined and this thesis would not have happened without you, so thank you.

My first contact at the IPS was Christoph Draxler. His warmth and openness towards me and all of his students as well as his belief in our capabilities makes him a wonderful role model. So I'll never forget how you, Christoph, came to Marburg to give a talk (and blew everyone's mind), but also ended up paving my way to Munich. Thank you.

The girls on the fourth floor (little socio-phonetics joke) were the best office mates I could have wished for: Rosa Franzke, Pia Greca, Mona Franke, Ramona

Schreier and her dog Leo. I have really enjoyed our coffee breaks, having four knowledgeable and kind experts in the room to discuss questions, and sharing the highs and lows of doing a PhD with you. Thanks also to my new office mate, Vanessa Reichel, who proof-read parts of this thesis.

I am also very grateful to everyone who joined the Young Researchers Kolloquium over the years, especially Jasmin Rimpler, Esther Kunay, Conceição Cunha, Andrea García Covelo, Lia Bučar Shigemori, and Sishi Liao. All of you kept me sane when we had to work from home during the pandemic and our meetings made me feel like part of a really lovely, amazing, and supportive community.

There are two people I need to mention, because they are the glue that keeps the IPS together: Klaus Jänsch and Ulrike Vallender-Kalus. Klaus, thank you for all of the user permissions you have given me over the years and for responding so quickly to any computer-related crisis. Frau Vallender-Kalus, thank you for helping me with such patience and competence whenever I had failed to fill out some paper work or needed administrative advice.

The IPS has become my second home since I first joined as a student in 2016. The amount of expertise at our institute is incredible and everyone so generously shares their knowledge. I would like to thank everyone at the IPS for creating such a productive, innovative, and friendly atmosphere. It has been a great joy to learn from and work with you.

I'd also like to thank the people outside the IPS with whom I had the pleasure to work: Ander Egurtzegi, Adèle Jatteau, Miša Hejná, Nicola Klingler, Michael Pucher, and Nick Henriksen. I am indebted to Hanna Ruch for providing me with an extensive database of Andalusian Spanish that has been fun to work on. This thesis was supported by European Research Council Grant No. 742289 "Human interaction and the evolution of spoken accent" (2017-2022) awarded to Prof. Jonathan Harrington.

Finally, and from the bottom of my heart, I would like to thank my family and friends for supporting me and believing in me even when I didn't.

Contents

Zusammenfassung vii

Abstract xiii

Acknowledgements xix

1 Introduction 1
- 1.1 The Study of Sound Change . . . 1
- 1.2 Overview . . . 8

2 A Dynamic Analysis of Aspiration Phases in Andalusian Spanish 13
- 2.1 Introduction . . . 14
- 2.2 Relation between Pre- and Post-Aspiration . . . 21
  - 2.2.1 Method . . . 21
  - 2.2.2 Results . . . 30
  - 2.2.3 Discussion . . . 31
- 2.3 Influence of Speakers' Age and Region on Closure Phasing . . . 33
  - 2.3.1 Method . . . 34
  - 2.3.2 Results . . . 35
  - 2.3.3 Discussion . . . 36
- 2.4 General Discussion . . . 38

3 An Agent-Based Model of Sound Change: soundChangeR 45
- 3.1 Introduction . . . 46
- 3.2 Computational Implementation . . . 52
  - 3.2.1 Agents and Exemplars . . . 53
  - 3.2.2 Initialisation of Agents . . . 54
  - 3.2.3 Production . . . 55
  - 3.2.4 Perception . . . 56
  - 3.2.5 Phonological Level . . . 58
  - 3.2.6 Memory Management . . . 61

3.2.7 From Interactions to Change . . . 62
3.3 Core Mechanisms . . . 64
3.3.1 Flexible vs. Fixed Phonology . . . 65
3.3.2 Phonetic Stability and Change . . . 68
3.4 Discussion . . . 76

4 Modelling the Aspiration Change in Andalusian Spanish 91
4.1 Introduction . . . 92
4.2 Aspiration in /st/ and /t/ . . . 96
4.2.1 Method . . . 97
4.2.2 Results . . . 100
4.2.3 Discussion . . . 108
4.3 Agent-Based Simulation . . . 109
4.3.1 Method . . . 110
4.3.2 Results . . . 113
4.3.3 Discussion . . . 119
4.4 General Discussion . . . 121
4.4.1 Handling Outliers . . . 122
4.4.2 Modelling Phonologisation . . . 125

5 Conclusion 129
5.1 Summary . . . 129
5.2 Insights . . . 134
5.3 Directions for Future Research . . . 136

A Appendices to Chapter 2 143
A.1 Authorship Contribution Statement . . . 143
A.2 Word List . . . 143
A.3 Mathematical details on A_{pre} and A_{post} . . . 145
A.4 Effects of FPCA-based signal decomposition . . . 149
A.5 Variation explained by PC2 and PC3 . . . 152
A.6 Including duration in FPCA . . . 158
A.7 Role of Areas, Time Normalisation, and FPCA . . . 160

B Appendices to Chapter 3 **165**
B.1 Authorship Contribution Statement 165
B.2 Installation of soundChangeR 165
B.3 Vignette to soundChangeR 166
B.4 Calculating Sub-Phonemic Classes 190
B.4.1 GMM-based Clustering and Classification 190
B.4.2 Non-negative Matrix Factorisation 192
B.4.3 Derivation of Acoustic and Sub-Phonemic Representations 194
B.5 Memory Management 202
B.5.1 SMOTE 202
B.5.2 Application of SMOTE in the ABM 204
B.6 Production-Perception Feedback Loop 205
B.7 Implicit Assumptions, Proxies, and Simplifications 207

C Appendices to Chapter 4 **211**
C.1 Word List 211
C.2 Median PC Score Values for Figure 4.3 212
C.3 Median PC Score Values for Figure 4.4 212
C.4 Estimated Marginal Means 213

Bibliography **215**

List of Figures **249**

List of Tables **252**

1 Introduction

1.1 The Study of Sound Change

Humans have been interested in the mechanics of their spoken language for more than two millennia (Schubiger, 1970). However, the field of science that investigates spoken language in all of its many facets – phonetics – was established as recently as the mid 19th century (Pompino-Marschall, 2009). Since then, great advances have been made with respect to our understanding of speech production and perception, not least because the methods with which these immensely complex processes can be observed and examined have evolved rapidly: from traditional anatomical studies to real-time imaging, from purely impressionistic to empirical descriptions, from analogue to digital recordings, and from manual to computational analysis. One of the most astonishing discoveries that emerged from research over the last 150 years is that spoken language is abundantly variable, to the point where any produced speech sound is never exactly like another even if produced by the same speaker in quick succession. But spoken language does not only vary synchronically, i.e. at a distinct point in time, but also diachronically, i.e. it changes over time. Why, for example, does French have nasalised vowels (e.g. *main* /mɛ̃/, engl. hand) when related languages such as Spanish (*mano* /mano/) do not? Why is the word-initial consonant "k" produced in German (e.g. *Knoten* /knoːtən/), but not in English (*knot* /nɔt/)? And why does the speech of actors in old black and white films sound so different from that of contemporary speakers? It is this phenomenon – the diachronic change of sounds in spoken languages – which is the central topic of this thesis.

Sound changes are the result of the extraordinarily complex and still poorly understood interplay of forces within speakers, listeners, their societal structures, and their language. The synchronic variability of spoken language mentioned above is considered the raw material for sound changes and can be attributed to the articulatory processes executed by the speakers. Not only are vocal tracts and articulators unique (Allen et al., 2003; Beck, 2010; Fant, 1960;

Harrington, 2014; K. Johnson et al., 1993; Peterson & Barney, 1952; Zellou, 2017), but the speakers' articulation is also to a certain degree conditioned by the norms of their society with regard to e.g. gender, ethnic background, and social class (Babel et al., 2014; Eckert, 1989; Hall-Lew et al., 2010; Hay et al., 1999; K. Johnson, 2006; Labov, 1963, 1972). While it seems astounding that people can understand each other and process their interlocutor's message correctly most of the time given this enormous variability in speech production, it seems even more miraculous when considering that speech perception is just as individual and variable. Besides physiological idiosyncrasies, language processing is influenced by experience. It has been shown that listeners memorise perceived speech in great detail (Campeanu et al., 2014; Church & Schacter, 1994; Goldinger, 1996, 1998; Goldinger & Azuma, 2004; Palmeri et al., 1993; Sheffert & Fowler, 1995) and that exposure to ambiguous variants of sounds can lead to a shift of the listener's perceptual boundaries (Clarke-Davidson et al., 2008; Connine & Darnieder, 2009; Eisner & McQueen, 2006; Eisner & Mcqueen, 2005; Fenn et al., 2003; Goldstone, 1998; Kraljic & Samuel, 2005, 2006; Norris et al., 2003; Samuel & Kraljic, 2009; Zhang & Samuel, 2014). Thus, a listener's speech sound recognition and categorisation is shaped by their unique exposure to and experience with spoken language (K. Johnson, 1997; Pierrehumbert, 2001, 2003a, 2006; Yu, 2013; Yu & Zellou, 2019). Nevertheless, there is some regularity and universality that governs the "pool of synchronic variation" (J. J. Ohala, 1989) briefly described here. For instance, the kind of variability that results from articulatory processes is often directional, e.g. voiced plosives are more likely to become voiceless than vice versa due to aerodynamic constraints and vowels are more likely to centralise in unstressed than in stressed position (Garrett & Johnson, 2013; Sóskuthy, 2013). These synchronic biases are also reflected in phonological patterns that have developed diachronically as well as in language typology. For instance, word-final obstruents are often devoiced, e.g. German *Rad* /ʁaːt/ (engl. bike) but *Räder* /ʁɛːdɐ/ (engl. bikes) (J. J. Ohala, 1997); and if a language has only one set of plosives, it is much more likely to only have voiceless as compared to voiced ones (J. J. Ohala, 1983). If the biases that introduce directional variability are universal, however, why does sound change remain a rare

event? Why does it happen in one language under one set of circumstances, but not in another one or at another point in time? This puzzle, posited most prominently by Weinreich et al. (1968) as the actuation problem, underlines that sound change is not just multifactorial, but also stochastic and therefore impossible to predict in advance.

The stochasticity of sound change is often attributed to a failure in perception or speech processing on the side of the listener. In his seminal body of work, John Ohala (e.g. J. J. Ohala, 1981, 1989, 1993a, 1993b, 2012) claimed that sound changes originate from the listener's occasional failure to compensate for the effects of coarticulation. That is, most of the time the listener is able to reconstruct the intended sequence of sounds by factoring out the variability created by coarticulation or biases. But if the listener does not apply these corrective rules (hypocorrection) or over-applies them (hypercorrection), the listener-turned-speaker might produce the same sequence of sounds in a way that reflects the lack or excess of normalisation applied in perception. For example, if the listener does not compensate for the coarticulatory nasalisation of the vowel in /man/, they might perceive and eventually reproduce something like /mãn/ or even /mã/. On the other hand, a purposefully nasalised vowel which appears in adjacency to a nasal consonant may be hypercorrected and thus perceived and reproduced as oral. Thus, in Ohala's theory of the origin of sound change, the listener's misapplication of acquired compensation rules can lead to a mini sound change (J. J. Ohala, 2012). By contrast, Björn Lindblom's H&H theory focuses on the speakers' active role in sound change, claiming that they adapt to their interlocutors' informational needs by hyper- or hypo-articulating (Lindblom, 1988, 1990, 1998; Lindblom et al., 1995). When the message is low in informational density or can be derived easily from context or prior knowledge, it is more likely that the speaker hypo-articulates, resulting in a higher likelihood of phenomena such as undershoot, centralisation, and coarticulation. Under these circumstances, the listener is also more likely to direct their attention to the speech signal instead of the message, and might therefore notice, store, and reproduce biased or coarticulated speech sounds. That is, the origin of sound change according to Lindblom lies in the adaptation of articulatory precision by the speaker, but

depends also on the listener's choice between the 'what'- and the 'how'-mode (Lindblom et al., 1995). Yet another widely used theory of sound change is provided by William Labov (e.g. Labov, 1963, 1966, 1972, 1990, 2001). His influential work has examined the role of social factors such as gender, age, social class or background, and ethnicity on the spread of a new variant. In other words, whether or not a speaker adopts a new variant or propels a sound change is dependent on their social indices and network. An opposing point of view is taken by Peter Trudgill (e.g. Trudgill, 1986, 1999, 2004, 2008a, 2008b, 2011; Trudgill et al., 2000) who argued that sound change – at least with regard to the emergence of new dialects – is not a question of identity, but of numeric dominance. That is, if speakers from two or more dialects or languages come into contact, as has often been the case through colonialism and settlements, the linguistic outcome of that situation can be determined by taking into account which variant is spoken by the largest amount of speakers. Intermediate variants can emerge if the amount of speakers from two dialects in contact are approximately equal.

These, as well as many other theories of sound change (including but not limited to Beckman et al., 1992; Beddor, 2009; Bermúdez-Otero, 2020; Blevins & Wedel, 2009; Bybee, 2015; Harrington et al., 2012; Harrington, Kleber, Reubold, Schiel et al., 2019; Harrington & Stevens, 2014; Kerswill & Trudgill, 2005; Kirby, 2013; Martinet, 1952; Phillips, 1984; Solé, 2010; Sóskuthy et al., 2018; Yu, 2013) typically focus on a subset of circumstances and factors that are associated either to the emergence or to the propagation of sound change. Although some models in recent years (Baker et al., 2011; Kirby & Sonderegger, 2015; Sóskuthy, 2013; Stevens & Harrington, 2014) aimed to contribute to a solution of the actuation problem formulated by Weinreich et al. (1968), a holistic model of sound change remains yet to be developed. Based on the literature review, we can devise some desiderata that should be fulfilled by such a model:

(a) The model should predict that sound changes are rare.

(b) The model should show which circumstances (including intra- and extra-linguistic factors) can trigger specific sound changes.

(c) The model should show how novel variants of sounds spread through a community of speakers.

(d) The model should account for a variety of diverse sound changes that have been observed in linguistically unrelated languages.

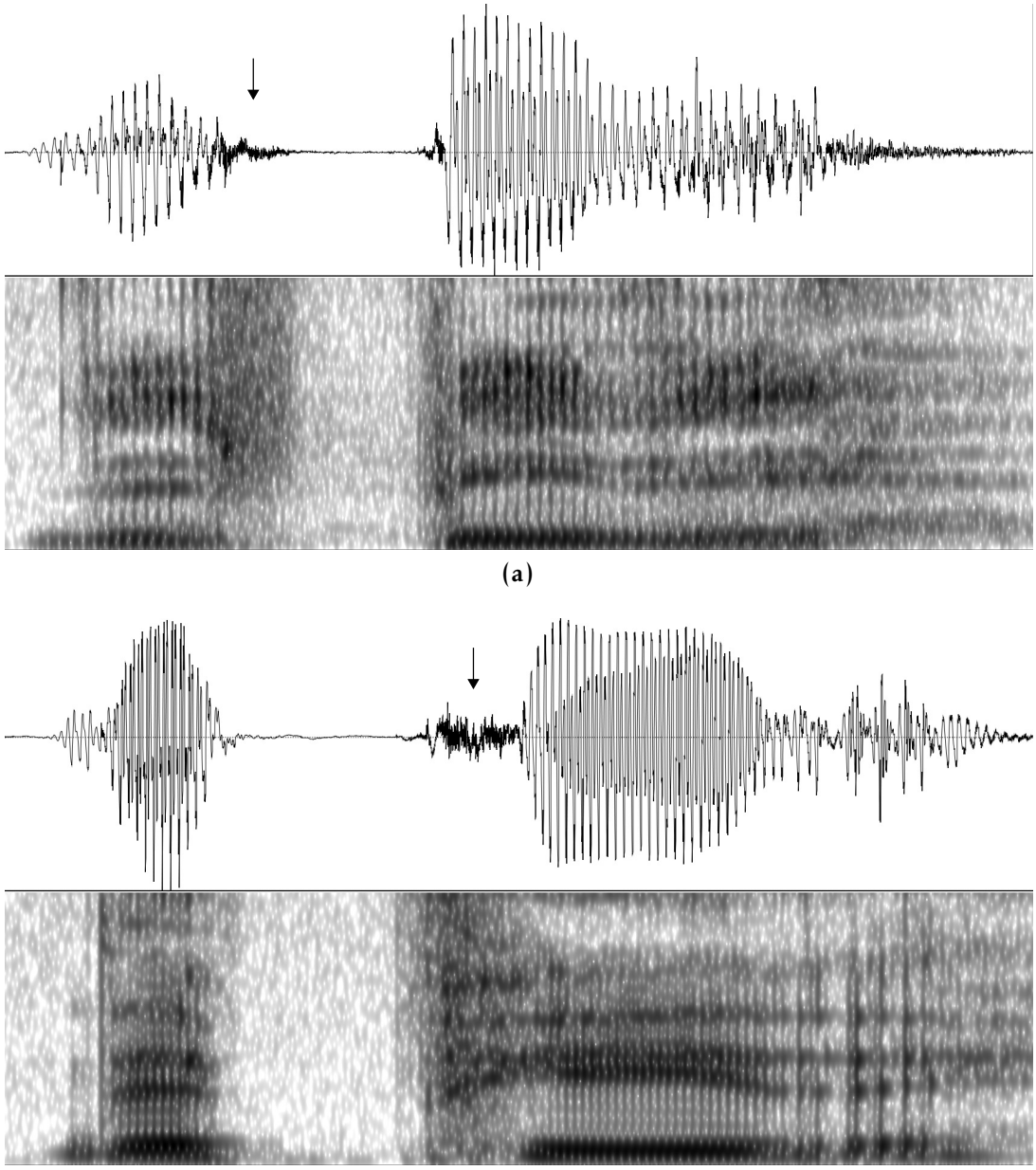
Even though the previously mentioned studies have enhanced our understanding of sound changes in many ways, developing this model remains an immensely complicated endeavour. Thankfully, the steady increase of computational power in the past two decades has enabled researchers to find new approaches to solving complex problems, one of which is agent-based modelling.

Agent-based models (ABMs) have found wide recognition as a helpful technique in many other scientific fields. Most prominently at the time of writing, ABMs were used to predict the spread of the Sars-Cov2 virus during the COVID-19 pandemic while taking into account vaccination rates, immunity levels after a vaccination, information on the infectiousness of novel strains of the virus, data on the movements of people provided by telecommunication companies, and many other factors (Castiglione et al., 2021; Gomez et al., 2021; Hackl & Dubernet, 2019; Keskinocak et al., 2020; Schlüter et al., 2021). Simulations like these can help inform political decisions in order to limit the spread of the virus. Another example of an agent-based model with political implications is the one by Hassani-Mahmooei and Parris (2012) which investigates internal migration movements in Bangladesh as a result of extreme weather events caused by climate change, poverty, and high population density. Similarly, Tonn and Guikema (2018) investigate how individual behaviour, technical interventions, and harm reduction efforts through policy influence a community's flooding risk. In the social sciences, applications of ABMs range from studies about the connection between police presence and crime levels (Wise & Cheng, 2016) to investigations of the factors that influence an election outcome (Laver & Sergenti, 2011). All of these ABMs have in common that the observed entity changes over time as a result of the interplay between factors inherent to the interacting individuals, their environment, or the entity itself.

These computational models are therefore also ideally suited for modelling changes of speech sounds.

The first of the two main aims of this thesis is to program, describe, and demonstrate the mechanisms of an ABM of sound change which can be used to gain insights that will eventually contribute to a holistic model of sound change. More specifically, the model presented here shows how sound change can emerge from the stochastic interactions between phonetically heterogeneous agents. The agents' speech production and perception is shaped by their exposure to diverse speech input and the agents can derive and reorganise phonological information from their own stored traces of speech. This model therefore provides an artificial world and controlled environment in which the role of and interplay between cognitive, social, and linguistic factors in sound change can be explored. Chapter 3 was written in pursuit of this first aim, with section 4.3 showing a use case of the ABM.

The second main objective of this thesis is to provide a dynamic analysis of a sound change in progress in Andalusian Spanish. This variety of Spanish is characterised by a debuccalisation or lenition of /s/, especially before voiceless plosives. An example of this can be seen in Figure 1.1a which shows the waveform and sonagram of the word *despide* (engl. she/he fires) produced by an Andalusian Spanish speaker. The arrow indicates the occurrence of friction noise right before the closure in /p/, i.e. the plosive is pre-aspirated. Pre-aspiration is considered a relatively rare feature in the world's languages (Bladon, 1986; Gilbert, 2023b; Silverman, 2003) and also in Andalusian Spanish it is currently in the process of being replaced by post-aspiration. That is, instead of producing the word *despide* as /de^{h}pide/, it is produced as /dep^{h}ide/, as shown in Figure 1.1b. The aspiration phase occurs after the closure in this case. Phonologically, however, Andalusian Spanish does not have post-aspirated voiceless plosives, e.g. the plosives in the word *pata* (engl. paw) are both unaspirated and remain unaffected by the change in /sp, st, sk/. The synchronic basis of the sound change by which pre- gives way to post-aspiration is believed to be a resynchronisation of articulatory gestures at faster speech rates (Parrell, 2012). More specifically, the way that the glottal opening is aligned with the oral closure determines whether a voiceless plosive is unaspirated (both gestures



(a)

(b)

Figure 1.1: Waveform and spectrogram of the word *despide* (engl. she/he fires) by **(a)** an older East Andalusian speaker who produced the /sp/ cluster with pre-aspiration, and **(b)** by a younger West Andalusian speaker who produced the cluster with post-aspiration. Arrows indicate the position of the aspiration phases. The spectrogram range goes up to 8 kHz.

must be synchronised and of the same duration), pre-aspirated (the closure is formed after the glottis has been opened), or post-aspirated (the closure is released while the glottis is still open). That is, according to this account from articulatory phonology (Browman & Goldstein, 1989, 1992; Goldstein & Browman, 1986) pre- and post-aspiration are inversely related through the timing of the closure with respect to the onset of the voiceless interval. So far, however, this sound change has only been investigated by measuring the duration of the aspiration phases. While duration measurements can certainly document a decrease in pre- and an increase in post-aspiration, this kind of static analysis cannot deliver any insights about the gestural realignment that may be at play in Andalusian Spanish. Moreover, duration measurements are based on an artificial and superimposed segmentation of the speech signal, which can be quite arbitrary and unreliable (Fowler & Smith, 1986). Indeed, sound changes arise from the dynamics of speech, i.e. from the imprecision of articulatory gestures and the resulting overlap and mutual interference of speech sounds. It is therefore vitally important to use methods in the investigation of sound changes that mirror the dynamic nature of spoken language. The second aim of this thesis is therefore not just to dynamically analyse the change from pre- to post-aspiration in Andalusian Spanish, but to show that it is feasible to capture the alignment of the articulatory gestures without having to collect physiological data. This aim was pursued in chapter 2 and section 4.2.

1.2 Overview

This thesis consists of three main chapters. Chapter 2 is concerned with a phonetic analysis of the change from pre- to post-aspirated /sp, st, sk/ in Andalusian Spanish that uses time-varying acoustic signals instead of a more traditional segmental approach. For 48 speakers of Andalusian Spanish, which are equally distributed across two age and two regional groups, the voicing probability (VP) and the high-frequency energy (HF) are measured between the two vowels on either side of the aspirated clusters. VP represents the glottal gesture, i.e. voicing or voicelessness, while HF represents the degree of closure as well as aspiration noise. The alignment of these signals indicates

the presence and location of aspiration around the closure since aspiration is characterised by voicelessness, friction noise, and lack of closure. These two signals are used as input for Functional Principal Component Analysis (FPCA), a technique that returns the main modes of variation in the signals' shapes (Principal Components, PCs). The analysis shows that the first PC captured the phasing of the closure in relation to the voiceless interval. This indicates that pre- and post-aspiration are inversely related to each other through the timing of the closure. Moreover, it is shown for the first time that the trade-off between pre- and post-aspiration is also signalled by the amplitude of the aspiration noise. Thus, this dynamic analysis is not only an appropriate investigation of the gestural realignment hypothesis, but also delivers new insights on the change towards post-aspiration. The first PC additionally distinguished between the speaker groups: younger speakers are shown to produce /sp, st, sk/ with more post-aspiration than older speakers who predominantly use pre-aspiration; the sound change is also further advanced in speakers from West than East Andalusia. These results establish a link between the synchronic basis (i.e. the resynchronisation of articulatory gestures) and a sound change in progress by means of the apparent-time approach. The conclusion from this study is that a resynchronisation of the closure relative to the voiceless interval is causally related to this sound change in progress by which pre-aspirated clusters of /s/ plus voiceless plosive come to be post-aspirated. Crucially, these results could only be achieved by means of this dynamic method and by using production data from a variety of speakers.

Chapter 3 is about the mechanisms and processes of a cognitively-inspired, computational agent-based model (ABM) which was implemented as an R package called `soundChangeR`. Agents in this model are representations of human individuals, i.e. they are initialised with production data from real speakers and are capable of producing and perceiving exemplars. Exemplars are parameterised traces of speech that are stored in the agents' memories together with the word in which the sound was uttered. Exemplars and words are linked via phonological classes which are either pre-determined by the user or computed by means of two machine learning algorithms. Simulations showed that the second option to derive phonological knowledge has

advantages over the first option because the machine learning algorithms were capable of identifying reasonable sub-phonemic classes in both systematically and randomly distributed (artificially generated) data. Moreover, this flexible phonology module is supported by usage- and exemplar-based theories of language which claim that phonological classes are abstractions over clouds of exemplars which are different from speaker to speaker because of their individual language exposure. Using another set of artificially created data, it was then demonstrated that the perceptual constraints greatly impact the simulation outcome. In `soundChangeR`, the agent listener evaluates whether or not to memorise a perceived exemplar by means of either one or both of the following two criteria: The relative criterion determines whether the exemplar is closer to its intended than to all other (sub-)phonological categories and thus tests the exemplar's discriminability. This criterion penalises acoustic ambiguity (i.e. ambiguous exemplars are not memorised) and can therefore maintain phonological contrasts. The absolute criterion, on the other hand, tests the exemplar's typicality by determining whether the exemplar is close enough in terms of its Mahalanobis distance to the intended (sub-)phonological class. The application of this criterion results in the reinforcement of phonetic biases because broad and skewed phonological classes are more likely to incorporate new exemplars than narrow ones. All in all, sound change in this ABM can emerge from the stochastic interactions between agents as a result of their production-perception loop and organisation of phonological information. The model's properties are discussed with respect to a variety of other computational models of sound change and some possible extensions to the current architecture are proposed.

Chapter 4 combines the dynamic analysis technique from chapter 2 and the agent-based model presented in chapter 3 to simulate the change from pre- towards post-aspiration in Andalusian Spanish clusters. In the first part, a dataset is composed of words containing either /st/ or /t/, the latter of which is considered to be phonologically unaspirated. Thus, /t/ posits a phonological contrast to /st/ which should not be affected by the sound change towards post-aspiration and can therefore serve as a sanity check in the simulation. The analysis showed that PC1 and, more clearly, PC4 separate /st/ from

/t/, i.e. the cluster is characterised by a longer voiceless interval, more high-frequency energy, and therefore more overall aspiration. According to this parameterisation, /t/ is produced with a small amount of post-aspiration which does not match its phonological description, but is also not unexpected given that the release of the oral closure lets air escape with some force for a few milliseconds before the vocal folds can start swinging again. PC2 captures the location of the aspiration phase in /st/, i.e. whether the cluster is pre- or post-aspiration (or a mixture of both). With the aim of modelling the change by which pre- gives way to post-aspiration, the PC scores extracted in the first part of chapter 4 are used as input to a simulation with `soundChangeR` in the second part. The agents in this simulation represent older and younger Andalusian Spanish speakers who exchange exemplars of /st/ and /t/. These two canonical phonemes are correctly identified and maintained by the flexible phonology module in `soundChangeR`, as expected. On the acoustic level, the agents produce /t/ the same way after as compared to before the interactions, but younger and older agents converge towards a common variant of /st/ that is characterised by both pre- and post-aspiration. This result can be attributed to the lack of skew in the older agents' PC scores towards those of the younger agents. Skew and orientation of phonemic classes in combination with selective memorisation are a key mechanism in `soundChangeR` which can trigger acoustic changes, as shown in section 3.3.2. Two further possible reasons for the ABM's failure to accurately model the change in Andalusian Spanish are discussed in the final part of chapter 4. The first is `soundChangeR`'s general bias against extreme exemplars which might pose a problem given that such outliers are particularly salient and might play an important role in sound changes. The second is that `soundChangeR` cannot model the re-weighting of perceptual cues which is considered an important part of sound changes which involve phonologisation.

The insights gained from the three main chapters are summarised in chapter 5 which also provides ideas and proposals for further research in this field.

2 | A Dynamic Analysis of Aspiration Phases in Andalusian Spanish

Abstract

In Andalusian Spanish, there is a well-documented sound change in which pre- has become post-aspiration in sequences of /s/ followed by voiceless stops. Here we investigate acoustically its synchronic basis across two age groups and two different regions of Andalusia that differ in the degree to which the sound change has advanced. For this purpose, Functional Principal Component Analysis (FPCA) was applied to the probability of voicing and to the degree of closure that had been estimated from the speech signal extending between the two vowels on either side of the aspirated cluster. The first principal component derived from FPCA was mostly associated with changes to the timing of the closure. Earlier closures were characteristic of both younger and West Andalusian speakers and of alveolar stops. In the signals parameterised by the first PC score, post- and pre-aspiration were found to be acoustically inversely related to each other and predictable from closure timing. The general conclusion is that the sound change by which pre- evolves into post-aspiration is a derivative of resynchronising the closure relative to the voiceless interval that emerges after decomposing speech signals varying over a wide range of speakers into principal components of variation.

This chapter was published as Cronenberg et al. (2020) and is printed here with permission from Elsevier. The layout has been adapted to fit the general layout of this thesis. See Appendix A.1 for the authorship contribution statement. Footnotes 1 and 2 were added to answer points raised by the reviewers of this thesis.

2.1 Introduction

The southern varieties of Spanish are derived from 13th century Castilian Spanish, when groups of speakers from different, yet mutually intelligible dialects came into contact with each other during the *Reconquista* (Villena-Ponsoda, 2008). The resulting dialects, i.e. the regiolects of Andalusia, Extremadura, the Canary Islands, and – because of emigration from Andalusia to South America – some American varieties of Spanish, have since undergone many changes, among them the lenition of the voiceless alveolar sibilant /s/ (Canfield, 1981; Villena-Ponsoda, 2008). This sound change, that dates back to the beginning of the 18th century according to Mondéjar Cumpián (2001) and Terrell (1980), manifests itself as /s/-debuccalisation and affects the sibilant in a wide range of positions: word-medially before consonants as in *este* /e^{h}te/ (engl. this), word-finally before consonants as in *las toman* /la^{h}toman/ (engl. they take them), word-finally before vowels as in *las alas* /la^{h}ala/ (engl. the wings), and in some cases even word-initially as in *sí* /^{h}i/ (engl. yes) (Momcilovic, 2009; Torreira, 2006). This type of reduction affects not only /s/ but also other fricatives in Andalusian Spanish. Mondéjar, for instance, shows that words like *ajo* (engl. garlic) have changed their place of articulation from uvular /χ/ to glottal /h/: /aχo/ > /a^{h}o/ (Mondéjar Cumpián, 2001). Also, the interdental fricative /θ/ can be lenited to /h/ in syllable-final position in the South of Spain, such as in *voz* /voθ/ > /vo^{h}/ (engl. voice). The debuccalisation of fricatives is a fairly common process in the languages of the world (Solé, 2010; Terrell, 1980). Consider for example the existence of prefixes such as *super-/hyper-*, *sex-/hex-*, *semi-/hemi-* in present-day English. Such alternations between /s/ and /h/ derive from borrowings from Latin, which did not lenite /s/ to /h/, and Classical Greek which has developed away from its Proto-Indo-European (PIE) roots and has undergone /s/-debuccalisation, e.g. PIE *septḿ* > Cl. Greek *hɛptá* (engl. seven) (J. J. Ohala, 1993b).

The debuccalisation of /s/ before voiceless plosives has led to the development of pre-aspirated plosives in Andalusian Spanish, e.g. in *esquina* /e^{h}kina/ (engl. corner) (Ruch & Harrington, 2014; Torreira, 2007). Romero (1994) links the historical development of pre-aspiration to the production of /s/ with a

laminal place of articulation in Andalusian as opposed to the post-alveolar, apical production of Castilian Spanish /s/. Romero (1994) also suggests that pre-aspiration may have been more likely to arise in Andalusian than in Castilian Spanish because, in comparison with an apical /s/, the gestures in producing a laminal /s/ are slower and less extensive as a result of which a laminal /s/ is more likely to be lenited.

The phonetic characteristics of the sibilant in /sC/ vary due to a number of factors such as stress, lexical frequency, and speech rate (Alvar, 1955; Parrell, 2012; Ruch & Peters, 2016; Terrell, 1980; Torreira, 2006; Villena-Ponsoda, 2008). There is mixed evidence about whether /s/-aspiration is accompanied by quality differences in the preceding vowel (see Herrero de Haro, 2017 for a review). Auditory impressions suggest that the preceding vowel has a more open quality if /s/ is aspirated or deleted (Navarro Tomás, 1938) leading to singular/plural distinctions in East Andalusian Spanish in final and principally mid vowels if the following /s/ is completely deleted, e.g. *paso* [ˈpaso] (engl. step) vs. *pasos* [ˈpasɔ] (engl. steps) (Hualde & Chitoran, 2016; Hualde & Sanders, 1995). Such differences in quality in mid vowels may also spread anticipatorily to the stem in a form of metaphony leading to distinctions such as *perro* [pe̞ro̞] (engl. dog) vs. *perros* [pe̝rɔ] (engl. dogs). However, further experimental data is needed to support these impressions (Torreira, 2007). Gerfen (2002) also provides some evidence of a greater duration of the consonant closure and of a correspondingly lesser duration of the vowel in pre-aspirated East Andalusian post-vocalic /sC/ words (e.g. *pasta* /pasta/ [pa^{h}tːa], engl. pasta) than in corresponding singleton /C/ words (e.g. *pata* /pata/ [pata], engl. paw), but only on the assumption that vowel duration was defined to extend from its acoustic onset to the offset of pre-aspiration.

The phonetic characteristics of /s/ in Andalusian Spanish also vary by age and regional origin of the speakers. Pre-aspiration in word-medial /sC/ clusters fairly consistently occurs in the speech of older speakers from Granada (East Andalusia) and may be accompanied by breathy voice during the preceding vowel or a longer closure duration (Gerfen, 2002; Torreira, 2006). Younger speakers from Seville (West Andalusia), on the other hand, were found to produce pre-consonantal /s/ either as pre-aspiration, or as post-aspiration,

e.g. /ek^{h}ina/, or as a combination of both, e.g. /e^{h}k^{h}ina/ (Ruch, 2013; Ruch & Harrington, 2014; Ruch & Peters, 2016; Torreira, 2012). These findings provide evidence for a sound change in progress from pre-aspirated to post-aspirated voiceless stops in Andalusian Spanish that, so far, has taken stronger hold in younger speakers from West Andalusia than in older speakers from East Andalusia.

Regular sound change is often directional due to the existence of a phonetic bias that promotes the change to work in one, but not the contrary direction (Harrington et al., 2018; Labov, 1994). In the case of Andalusian Spanish, it is a faster speech rate that favours a decrease in pre-aspiration and an increase in post-aspiration. Parrell (2012) has shown that the /st/ cluster in the word *pastándola* (engl. grazing it) exhibits more post-aspiration and less pre-aspiration in fast than in slow speech. Similarly, it has been found for Cuban Spanish that the educated societal class retains /s/ as a sibilant in 22% of all cases in semiformal speech, but only in 3% in fast and informal speech (Terrell, 1980). The weaker perceptual salience of pre-aspiration in comparison with post-aspiration may be an additional bias (Bladon, 1986; Ruch, 2018; Ruch & Harrington, 2014). In the production of post-aspirated, but not pre-aspirated stops, there is a build-up of air-pressure behind the closure which results in an acoustic burst (Fant, 1973) and a rapid modulation of the acoustic signal when the stop is released. The strong perceptual salience of a post-closure release is demonstrated in perception experiments in which listeners' judgements of place of articulation in heterorganic stop consonant clusters C_1C_2 in VC_1C_2V are swayed far more by C_2 than by C_1 (J. J. Ohala, 1990). Moreover, in a perceptual experiment by Ruch and Harrington (2014), listeners of Argentinian Spanish who typically produce /sC/ clusters with pre-aspiration and no post-aspiration (Aleza Izquierdo & Enguita Utrilla, 2002; Torreira, 2006) were more inclined to perceive *pasta* in a continuum synthesised between singleton *pata* (engl. paw) and *pasta* (engl. pasta) when the continuum was created with a long (27 ms) as opposed to short (13 ms) duration of post-aspiration. Thus, Argentinian Spanish listeners are nevertheless influenced by post-aspiration as a cue to the distinction between post-vocalic /st/ and post-vocalic /t/, even though they produce pre- and not post-aspirated stops

and even though (in contrast to Andalusian Spanish) there is no evidence of a sound change in progress in this variety in which pre- are becoming post-aspirated stops.

The synchronic basis for the sound change by which pre- has evolved into post-aspiration as proposed by Parrell (2012) is a resynchronisation of autonomous articulatory gestures that also typically occurs at faster rates of speech (cf. e.g. Beddor, 2009; Davidson, 2006). In this model, depicted in Figure 2.1, the oral tract constriction gesture (solid) for the stop closure shifts to be in phase with the glottal opening gesture (dashed) at faster rates of speech. A direct consequence of the shift from anti- to in-phase timing is a decrease in pre-aspiration and an increase in post-aspiration strength: that is, a by-product of this resynchronisation is that the articulatory durations of pre- and post-aspiration stand in an inverse relationship to each other.

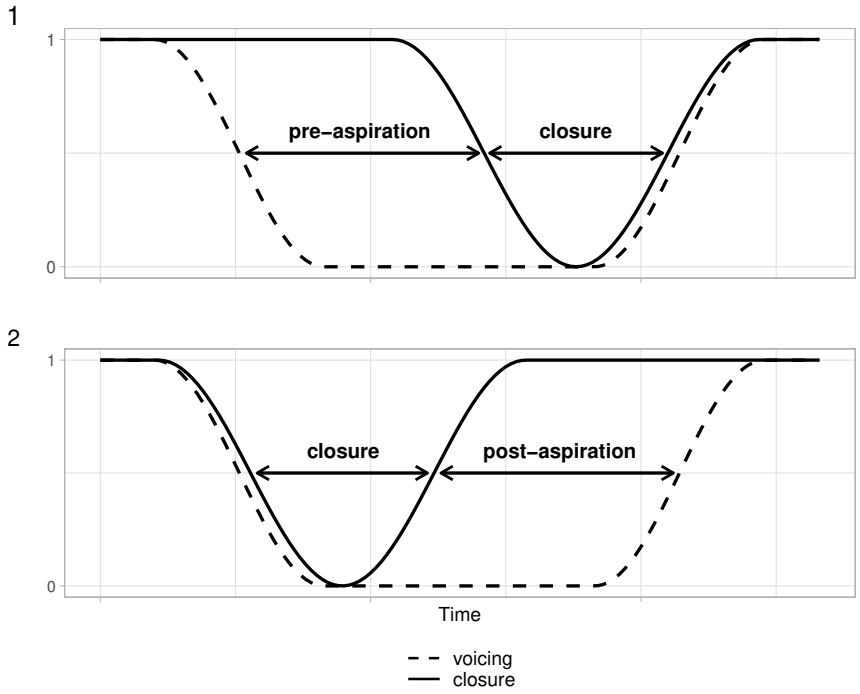


Figure 2.1: Idealised scheme of resynchronisation of the closure with the voiceless interval in Andalusian Spanish /s/-aspiration. The solid line is the glottal gesture, where low values stand for an open glottis and hence voicelessness, the dashed line is the oral constriction gesture of the voiceless plosive, where the minimum of the curve indicates maximal closure.

It is possible – although so far undemonstrated for aspiration in Andalusian Spanish – that this inverse articulatory relationship forms the basis of a perceptual trading relationship by which listeners parse aspiration from the speech signal but might be agnostic about its temporal location (about whether aspiration occurs before the closure, after the closure, or both; Ruch and Harrington, 2014). There is a potential analogy to a different type of sound change involving the phonologisation of nasalisation studied by Beddor and colleagues in recent years (Beddor, 2009, 2012; Beddor et al., 2018). Their model is informed by at least four related sets of studies from speech perception. Firstly, listeners are sensitive to anticipatory coarticulation in perceiving speech (Alfonso & Baer, 1982; Martin & Bunnell, 1982): listeners perceive a speech sound close to its articulatory onset, i.e. from the time at which there is coarticulatory evidence for the speech sound in the acoustic signal (Fowler, 1984, 2005). Secondly, listeners weight differently the multiple cues to speech sounds that originate from overlapping and coproduced speech gestures (Boersma et al., 2003; Clayards, 2018; Francis et al., 2000; Holt & Lotto, 2006; Idemaru et al., 2012). Thirdly, listeners have the capacity to re-weight cues (Boersma et al., 2003; Francis & Nusbaum, 2002; Harmon et al., 2019; Idemaru & Holt, 2011) as also shown by studies of perceptual learning (McQueen et al., 2006; Norris et al., 2003; Reinisch and Holt, 2013, see Samuel and Kraljic, 2009 for a review). Cue reweighting in perception has also recently been shown to carry over to speech production (Lehet & Holt, 2017). Fourthly, the cues can enter into a so-called trading relationship (Beddor, 2009; Best et al., 1981; Haggard et al., 1981; Kingston et al., 2008; Kirby, 2014b; Repp, 1982; Whalen et al., 1990) in which listeners can pay more attention to a secondary cue if the primary cue to a phonological contrast is compromised. In many types of sound change such as the development of contrastive vowel nasalisation, metaphony (Savoia & Maiden, 1997; Torres-Tamarit et al., 2016), and tonogenesis (Hagège & Haudricourt, 1978; Hombert et al., 1979), secondary cues that were initially brought about by coarticulation become primary while the primary cues are downweighted and typically completely disappear (thus leading to e.g. the development of present-day German Füße /fysə/ (engl. feet) from old High German /fotiz/ in which there is no trace left of /i/ which

initially caused via VCV coarticulation the fronting of the high back vowel in the first syllable; see Kiparsky, 2015; Penzl, 1949; Twaddell, 1938). This stage, in which the secondary cues become primary and primary cues have all but vanished is when the sound change is in the process of phonologisation (Bermúdez-Otero, 2015; Bermúdez-Otero & Trousdale, 2012; Hyman, 2013; Kiparsky, 2015; Ramsammy, 2015). Beddor showed that the precursor to phonologisation is the development of a perceptual trading relationship (Beddor, 2009, 2012). When the cues for a feature trade, then listeners perceive the feature but without necessarily associating it with the coarticulatory source or effect: thus, in American English *send*, they hear nasalisation somewhere in the rhyme without parsing the nasalisation with either of the rhyme's constituents (the vowel or the coda /n/).

The studies by Beddor and colleagues further suggest that the physiological basis of such a trading relationship is an inverse relationship between the coarticulatory source and effect. As far as nasalisation is concerned, this is modelled as the temporal sliding of a nasal gesture of constant articulatory duration into the gesture for the vowel (Beddor, 2009; Beddor et al., 2007): the more the two gestures overlap, the greater the extent of vowel nasalisation and the smaller the degree of articulatory prominence of the post-vocalic nasal.

The purpose of the present study is to determine from an acoustic analysis whether there is any evidence of this type of gestural sliding in the production of /s/-aspiration in Andalusian Spanish that could in turn form the basis of a perceptual trading relationship as proposed by Beddor. There are evidently similarities in the proposed physiological model that underlies both types of sound change. Thus, in both Beddor's model of the production of nasalisation and in that proposed by Parrell (2012) for /s/-aspiration, the consequence of sliding the gesture of constant duration in the direction that eventually leads to the sound change is that, as one of the cues wanes (post-vocalic /n/; pre-aspiration) the other becomes more prominent (coarticulatory vowel nasalisation; post-aspiration).

One of the main aims of the present study is to analyse /s/-aspiration in Andalusian Spanish in order to determine whether there is any evidence for the type of model proposed by Parrell (2012) in which pre- and post-

aspiration stand in an inverse relationship to each other as a consequence of the resynchronisation of articulatory gestures of relatively constant duration. Neither studies by Ruch and colleagues (Ruch, 2013; Ruch & Harrington, 2014; Ruch & Peters, 2016) nor Torreira (2007) have found much evidence in support of such a relationship. Moreover, Ruch and Harrington (2014) showed that the closure duration, far from being stable, was influenced by the age and regional origin of the speakers. On the other hand, none of these studies nor indeed Parrell's (Parrell, 2012) are necessarily appropriate tests of the model in Figure 2.1 because they are based on vertical segmentations of the speech signal (as defined in van der Kooij and van der Hulst, 2005, p. 167) into pre- and post-aspiration and a closure instead of modelling more directly how glottal and supraglottal gestures are aligned and potentially resynchronised. Segmentations are often unreliable and arbitrary given the dynamic nature of speech (Fowler & Smith, 1986), so it might be the case that a trading relationship can only be identified by considering a representation of speech that is a closer representation of the model in Figure 2.1. Moreover, the type of trading relationship that has been identified in perception in connection with sound changes in progress might come about only after a listener has experienced talkers that differ in their use of the acoustic cues that are traded in perception. For example, a listener might interpret perceptual equivalence between $\tilde{V}$ and VN only after having been exposed to many talkers that differ in the extent to which they nasalise a vowel in a VN context. If this is so, then a trading relationship might be related to variations regulated by dynamic relations between two cues that characterise an entire population of speakers.

The present study addresses these issues by computing two time-varying signals from acoustic data of Andalusian Spanish, one being a proxy for the glottal gesture, the other a proxy for the oral constriction gesture. According to articulatory phonology (Browman & Goldstein, 1986, 1992), the alignment of these two gestures with respect to each other is responsible for the presence or absence of aspiration around the voiceless plosive C in sequences like V_1sCV_2. A central concern in section 2.2 will be to test whether there is any evidence across the population of speakers that pre- and post-aspiration are negatively

related through the phasing of the oral closure. The aim in section 2.3 is to investigate whether age and regional origin of the speakers condition the timing of the oral closure with reference to the voiceless interval. If so, this would provide a link between the synchronic model in Figure 2.1 and a sound change in progress by which pre-aspiration wanes and post-aspiration increases in Andalusian Spanish /sC/ clusters. In the last part of this paper, we will discuss how the dynamic analysis of speech can be beneficial to the study of sound change and outline a new cognitive-computational model of sound change which is based both on principles from articulatory phonology and episodic models of speech.

2.2 Relation between Pre- and Post-Aspiration

There were two main issues of concern here. The first was to build a general model of the synchronisation of the closure with the glottal opening (Figure 2.1) based on the entire database of V_1sCV_2 sequences across speakers of different age groups and regions. Functional Principal Component Analysis (FPCA) was used for this purpose since it can model two (or more) time-varying signals simultaneously (Gubian et al., 2019; Gubian et al., 2015). The second was to test in this general model the extent to which pre- and post-aspiration are predictable from closure phasing.

2.2.1 Method

2.2.1.1 Speakers and Materials

The data analysed in this study was taken from a larger speech database collected by the fourth author of this paper (Ruch, 2013). Parts of this database have been analysed elsewhere (Ruch, 2013; Ruch & Harrington, 2014; Ruch & Peters, 2016). The focus of the present study was on 48 speakers of Andalusian Spanish who produced words containing /sC/ clusters in a V_1sCV_2 context (C = /p, t, k/). The speakers were equally divided between two age groups (younger: 20-36 years; older: 55-79 years) and two regions (East and West):

thus, there were 12 speakers for each of the four possible age group × region combinations. The clusters occurred in the words listed in Appendix A.2. The words in the available corpus had been constructed as far as possible to include /s/-aspiration clusters in several mostly high frequency real words for three places of articulation /sp, st, sk/ combined with one of the three vowel types /i, u, a/. The majority (just over 80%) of words had a paroxytonic lexical stress pattern with primary lexical stress on the penultimate syllable (e.g. *espía* /esˈpia/). Other stress patterns (e.g. *pasta* /ˈpasta/) or non-words (e.g. *bestiando*) were used when there were an insufficient number of real words for these place × vowel combinations.

The productions were elicited using a prompt in which each word appeared individually or in a short sentence on a computer monitor (see Ruch and Harrington, 2014 for further details of the recording procedure). There were up to three repetitions per speaker per word, thus giving a potential maximum of 48 (speakers) × 52 (words) × 3 (repetitions) = 7488 word tokens. However, not all speakers produced three repetitions. Tokens which included errors of production (in particular the production of the wrong target word or of a false start) as well as any productions of target words with standard Spanish /s/ in the /sC/ cluster were removed from further consideration following the procedure in Ruch and Harrington (2014). This left 6393 word tokens. Furthermore, 446 productions in which the voicing probability peak in V_1 and/or V_2 was too low to indicate voicing were also discarded since it was then not possible to identify reliably the existence of pre- or post-aspiration. Another 82 tokens were excluded because the PEFAC algorithm (see section 2.2.1.2) erroneously computed the voicing probability peaks to occur during the closure, not during the neighbouring vowels. A further 18 tokens were excluded from the analysis because the target words were not fully recorded, i.e. the recording was clipped at the end. Two tokens had to be excluded because of technical issues during the automatic speech processing. The final count of the analysed word tokens

was 5845, distributed across 48 speakers and 52 word types that included 20, 17, 15 word types containing /st, sp, sk/ clusters respectively.[1]

2.2.1.2 Acoustic Parameters

The synchronisation of the closure with the glottal opening was estimated acoustically by means of two separate parameterisations of the acoustic signal. The first of these, designed to model the glottal opening, was the voicing probability (henceforth VP) that was computed by applying the PEFAC algorithm (Gonzalez & Brookes, 2014) to the original audio files with a 5 ms frame shift and otherwise default settings. The second, which was used to model the supralaryngeal closure, was derived from the high frequency energy (henceforth HF) in the speech signal. This was obtained by double-differencing the audio signal to give 12 dB boost per octave (i.e. pre-emphasis) resulting in sharper transitions between fricated and closure phases of the signal (Harrington & Cassidy, 1999). These double-differenced signals were then high-pass filtered at 3 kHz, from which the logarithmic root mean square energy was computed with a window length of 20 ms and a frame shift of 5 ms. The resulting signal was smoothed with a 20 Hz Butterworth low-pass filter (Butterworth, 1930), then normalised such that 0 dB was set to the minimum of this signal during the supralaryngeal closure (point *M* in Figure 2.2). An example of the two resulting signals, HF and VP, for one utterance of the word *estado* (engl. state) by an older West Andalusian speaker can be seen in panels 2 and 3 of Figure 2.2, respectively. The interval that was processed within V_1sCV_2 was

[1] A reviewer of this thesis suggested to go into more details about the relatively high percentage of excluded data and the effects this might have on the analysis. The primary reason for excluding data is related to eliciting dialectal speech. The formal recording setting, the non-native experimenter, and the reading task (note that Andalusian Spanish does not have a written standard) can make it difficult for participants to produce their dialect (Bailey, 2017). In addition, older speakers can sometimes struggle to read aloud for the duration of the experiment, leading to less repetitions and clipped audio files. Given that the aim of this study was to analyse a regional variant of a sound, it was important to include only non-Standard productions so as to ensure that the results reflect the dialectal characteristics under investigation. An effect of /s/-aspiration that could not be studied using the methodology presented here was the devoicing of the vowels surrounding the aspirated cluster. For future studies, it might thus be interesting to analyse the 446 tokens which were excluded because the voicing probability was too low in V_1 and/or V_2.

defined as extending between points A and B in Figure 2.1. Point A is the time at which the voicing probability first attains its maximum value in V_1 working backwards in time from the time of maximum closure, M. Point B is the time at which the voicing probability first attains its maximum value in V_2 working forwards in time from M. The other parts of the signal from the acoustic onset of V_1 to A and from B to the acoustic offset of V_2 were excluded from the analysis.[2]

2.2.1.3 Data Analysis

Functional Principal Component Analysis (FPCA) (Gubian et al., 2015; J. O. Ramsay & Silverman, 2010) was used in order to find the main dimensions of variation in the $N = 5845$ pairs of $HF_i(t)$ and $VP_i(t)$ curves, $i = 1,\ldots,N$. There were three pre-processing steps prior to applying FPCA. Firstly, in order to obtain a functional representation from the time-sampled curves, standard smooth interpolation techniques were applied using B-splines as function basis (see Gubian et al., 2015 and J. O. Ramsay and Silverman, 2010 for details). Secondly, the signals were linearly time-normalised between the times of the voicing probability maxima (time points A and B in Figure 2.2; see Appendix A.6 for details of the effect of time normalisation). Thirdly, the HF signals were also amplitude re-scaled by dividing each curve by the 75% quantile of all HF values. This was done to ensure that HF and VP signals both spanned approximately the same range between 0 and 1.[3] As a consequence, HF and VP were a closer representation of the model in Figure 2.1. This re-

[2] A reviewer of this thesis noted that the relation between the articulatory gestures and the chosen acoustic proxies could be discussed in more detail. While articulatory data is more resource-intensive to collect and process than acoustic data, it might also allow for a more fine-detailed insight into the alignment of the gestures involved in the production of (aspirated) voiceless plosives. However, the main effects of these gestures – namely, the presence or absence of voicing, closure, and aspiration noise – can be measured acoustically to some level of accuracy, as shown in Figs. 2.2 and 2.3 and as stated by Fowler (2005) and Fowler and Brown (2000). Of particular interest is the synchronisation of the point of full oral closure with the peak glottal opening, although as this study shows the dynamics of these movements are equally important.

[3] The empirical choice of the 75% quantile as a normalising factor was justified by the observation that the distribution of HF values has a long right tail; thus an exact normalisation based on the maximum would have compressed the HF signals to the extent that the VP signal would dominate (VP > HF) most of the time.

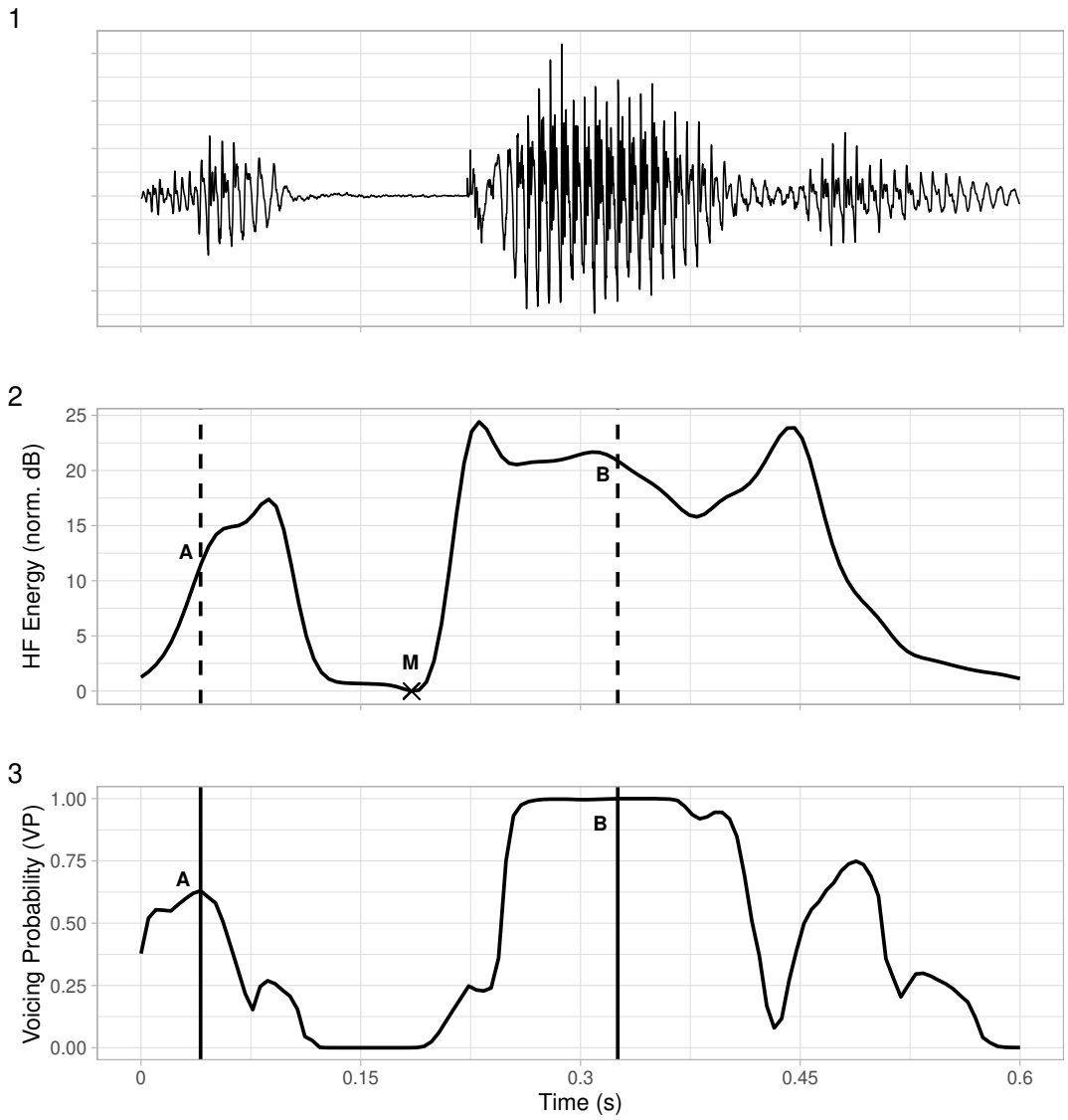


Figure 2.2: From upper to lower panels: Waveform, HF (amplitude-normalised high frequency energy), and VP (voicing probability) signals over time for a production of *estado* (engl. state) by an older West Andalusian speaker. *A* & *B* are the voicing probability maxima in V_1 and V_2, respectively. *M* is the point of lowest amplitude during the closure. The interval that was analysed extended between *A* and *B*.

scaling also prevented the principal components from being unduly influenced by large amplitude values of one of the signals with respect to the other.

The FPCA parameterisation was expressed by the following pair of equations applied between time-normalised values 0 and 1 (e.g. points A and B in Figure 2.2):

$$HF_i(t) \approx \mu_{HF}(t) + \sum_{k=1}^{K} s_{k,i} \cdot PCk_{HF}(t) \tag{2.1a}$$

$$VP_i(t) \approx \mu_{VP}(t) + \sum_{k=1}^{K} s_{k,i} \cdot PCk_{VP}(t) \tag{2.1b}$$

where $\mu_{HF}(t)$ and $\mu_{VP}(t)$ are the mean signals, e.g. $\mu_{HF}(t) = \frac{1}{N}\sum_i HF_i(t)$, the functions $PCk_{VP}(t)$ and $PCk_{HF}(t)$ are K pairs of Principal Components (PCs), $k = 1,\ldots,K$, which are based on the entire data set, and $s_{k,i}$ are weights or *scores*, which modulate each PCk differently for each signal pair $(HF_i(t), VP_i(t))$. Formally, Eq. (2.1) follow the same structure of ordinary PCA in which any input signal is approximately decomposed into a linear combination of K PCs added to the mean. The main difference in comparison with PCA is that in FPCA the input, mean, and PCs are functions of time as opposed to vectors of real numbers. Crucially, the linear combination expressed by PC scores modulates the PCs for both dimensions *together*, i.e. $s_{1,i}, s_{2,i}, \ldots, s_{k,i}$ are the same in Eq. (2.1a) and (2.1b), which is essential for capturing systematic co-variations across dimensions. Using the R package `fda`,[4] we computed the first $K = 3$ PCs for the set of 5845 smoothed curve pairs of $HF(t)$ and $VP(t)$ which explained 31.5%, 24.3%, and 14.5% of the variance in the curve shapes respectively (70.3% combined). Only the first PC is considered in the remainder of this paper since the second and third added very little to explaining phonetic variation in the data that was relevant to the sound change at hand (see Appendix A.5 for more details on PC2 and PC3).

The resulting PCs and PC scores can be used to reconstruct individual pairs of signals. The reconstructed curves tend towards an ever closer approximation to the original signals as the number of PCs that are used in the reconstruction

[4] Version 2.4.0 was used here. More recent versions (5.1.5.1 being the current one at the time of writing) can be used as well, provided that PC scores are summed across dimensions.

is increased (i.e. for increasing values of K in Eq. (2.1); see Figure 4 in Gubian et al., 2015 for an example). This operation is demonstrated in Figure 2.3 for one clearly post-aspirated and one clearly pre-aspirated instance of the word *despide* (engl. he/she/it fires) produced by two different speakers. The normalised time points 0 and 1 correspond to the beginning and end respectively of the sequence of interest V_1sCV_2 (cf. points A and B in Figure 2.2). Row 2 in this plot shows HF (solid) and VP (dashed) which were obtained as described in section 2.2.1.2. Row 3 instead shows the HF and VP signals which were reconstructed using only PC1 and the corresponding PC score s_1 (0.46 for the post-aspirated token, −0.48 for the pre-aspirated token) in Eq. (2.1). It is clear from Figure 2.3 that the reconstructed curves in row 3 are similar in shape, but smoother versions of the raw HF and VP signals in row 2 that were derived directly from the speech signal. Rows 2 and 3 of the left column of Figure 2.3 show the in-phase timing of the articulatory gestures represented by HF and VP, in which the closure of the plosive and the start of the voiceless interval are approximately synchronous, resulting in post-aspiration. By contrast, HF and VP in the right column of Figure 2.3 show an anti-phase timing in which the closure is delayed relative to the start of the voiceless interval, resulting in pre-aspiration. Recall that we expect there to be a closure when both HF and VP show very low values, i.e. when there is very little energy as well as hardly any or no voicing. Overall, the HF and VP signals matched the waveforms very well. While there can be some imprecisions due to erroneous VP values or residual noise in the HF signal, a visual inspection of several signal pairs showed that the curves represented quite consistently and accurately the pre- and post-aspiration phrases.

2.2.1.4 Estimating Aspiration

In order to quantify the extent of aspiration before and after the closure, the area A_{tot} enclosed between the two curves HF and VP for HF > VP was measured. The reason why A_{tot} is an appropriate measure for aspiration is that aspiration occurs whenever VP as the proxy for the glottal gesture is low (i.e. no voicing) and HF as the proxy for the oral constriction gesture is

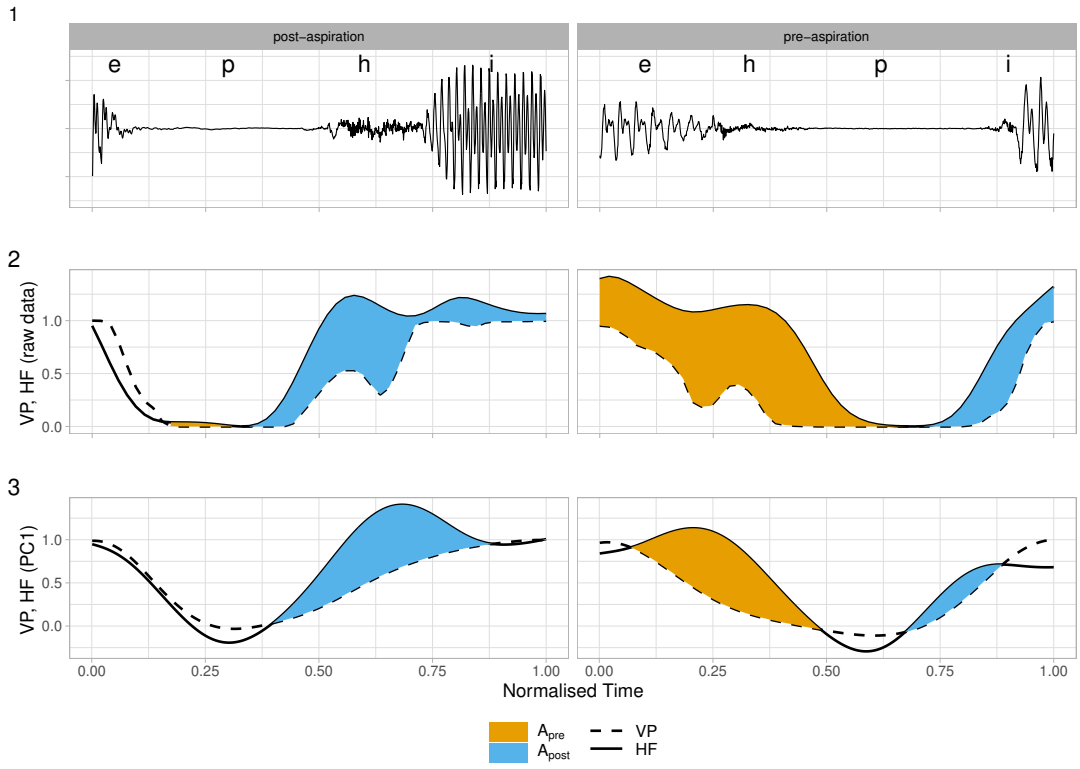


Figure 2.3: Example of a post-aspirated and a pre-aspirated token extending between points *A* and *B* (see Figure 2.2) in V_1sCV_2 in the word *despide* (engl. he/she/it fires). The post-aspirated token (left column) was produced by a young West Andalusian speaker, the pre-aspirated token (right column) was produced by an older East Andalusian speaker. The first row shows the waveforms, the second row shows HF (solid) and VP (dashed) calculated on the raw speech signals as described in section 2.2.1.2, the third row shows HF and VP as reconstructed based solely on PC1 using Eq. (2.1). The yellow areas are A_{pre} and the blue areas are A_{post} as in Eq. (2.2).

high (i.e. no closure), as suggested by Figure 2.1. The influence of pre- and post-aspiration that precede and follow the closure was respectively estimated from the areas A_{pre} and A_{post} defined in the same way as A_{tot} but spanning the normalised time intervals up to and beyond the minimum of HF, respectively (hence $A_{tot} = A_{pre} + A_{post}$). Formally:

$$A_{pre} = \int_{(HF>VP)\cap(0\leq t\leq t_M)} \Big(HF(t) - VP(t)\Big)\, dt \tag{2.2a}$$

$$A_{post} = \int_{(HF>VP)\cap(t_M\leq t\leq 1)} \Big(HF(t) - VP(t)\Big)\, dt \tag{2.2b}$$

where $t_M = \arg\min_t HF(t)$ (cf. point M in Figure 2.2, panel 2). The constraint HF > VP was introduced in order to exclude negative areas which, in contrast to the positive areas which we take as an estimate of aspiration strength, would not have any meaning in this context.

This method for inferring aspiration strength from area calculations is entirely independent of the functional data analysis, i.e. Eq. (2.2) does not specify how HF and VP were obtained. Examples of the area calculation both on raw (row 2) and reconstructed data (row 3) can be seen in Figure 2.3. That is, wherever HF > VP, A_{pre} (yellow) was computed as the area between the curves from normalised time point 0 to the point of maximal closure, and A_{post} (blue) was computed as the area between the curves from the point of maximal closure to time point 1. Figure 2.3 shows that A_{post} is large while A_{pre} is very small for the post-aspirated token on the left, whereas the opposite is true for the pre-aspirated token on the right. This was so both when HF and VP were obtained from the raw speech signals and when they were reconstructed using only PC1. It can also be seen that the areas partially extend into the vocalic parts of the segments. This is because aspiration (and in particular pre-aspiration) could often overlap with the vowel in a breathy voice production. This type of temporal overlap can be expressed by the area measurements proposed here, but is far more difficult to represent using durations extracted from vertical segmentations of the acoustic speech signals into the vowel and aspiration.[5]

[5] Appendix A.7 shows how the methods which were chosen for data analysis and the quantification of aspiration compare to various other methods, including the use of more conventional segmental durations.

2.2.2 Results

We consider firstly the variation in PC1 derived from an application of FPCA to the whole data set; and secondly an analysis of the how pre- and post-aspiration are related based on area calculations.

Figure 2.4 shows the relationship between quantitative changes in the first PC score and qualitative changes in the two signals HF and VP. The middle panel contains the mean signals $\mu_{HF}(t)$ (solid) and $\mu_{VP}(t)$ (dashed) across all input signals. These mean curves change, however, when s_1 is set to positive (right column) or negative values (left column), and all other scores to zero (recall Eq. (2.1)). Thus, the panels from left to right show how the shapes of the VP and HF signals are modified as the first PC score changes from negative to positive values. We chose these representative PC score values to be $\pm\sigma_{s_1}$, i.e. we added to, or subtracted from, each mean curve only the PC1 curve multiplied by 0.28, the standard deviation of the first PC score s_1.

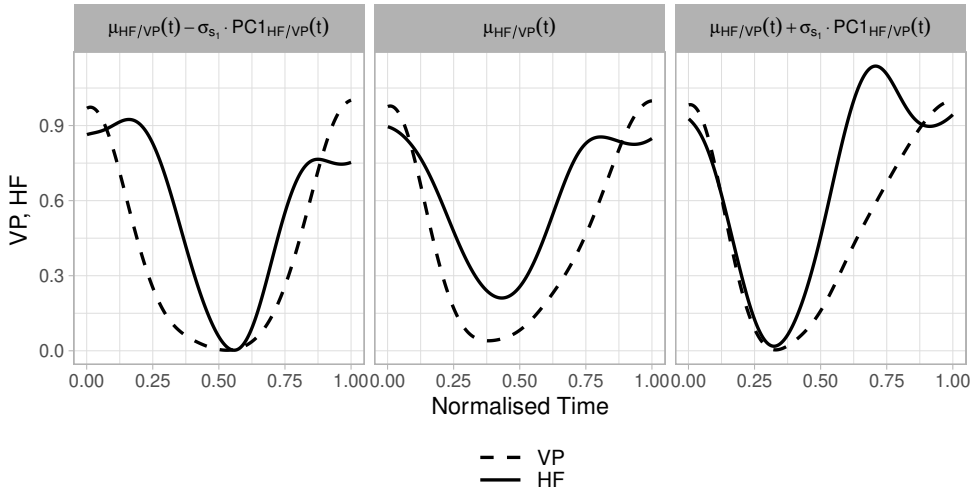


Figure 2.4: Variation expressed by PC1. The middle panel shows the mean curves $\mu_{HF}(t)$ and $\mu_{VP}(t)$, which were modified by adding to (right panel) or subtracting from (left panel) each mean curve the PC1 curve multiplied by the standard deviation of s_1. The exact formulae are given in the panel headings.

PC1 was closely related to the dynamic changes predicted by the model in Figure 2.1. A comparison of the left, central, and right panels of Figure 2.4 shows that PC1 modelled a phase shift relative to the VP signal of an HF-

minimum that corresponds to the instant when the maximal constriction of the vocal tract is attained. More specifically, variations from negative to positive s_1 resulted in a shift of the signal associated with the closure from late to early. Note, however, that it is not a necessary consequence of FPCA that the first PC describes a phase shift, nor that it coincides with the shape variation that is most relevant for the analysis it was employed for, as no prior information on the relevance of a phase shift was introduced as input to FPCA. The first PC simply explains (by definition) the largest amount of variance in any given data set. In the case reported here, it is coincidental that the most relevant kind of variation was captured by PC1 (the reader is referred to Appendix A.5 for an analysis of PC2 and PC3).

A test was then made of whether the closure phasing expressed by PC1 conditioned the extent of aspiration before or after the closure. To this end, HF and VP were derived from Eq. (2.1) using only PC1 (cf. Figure 2.4, or Figure 2.3, row 3) and the areas A_{pre}, A_{post}, and $A_{tot} = A_{pre} + A_{post}$ were computed using Eq. (2.2). Figure 2.5 shows these areas as a function of the first PC score s_1. The steep and opposite trends of the lines for $A_{pre}(s_1)$ (yellow) and $A_{post}(s_1)$ (blue) indicate that pre-aspiration and post-aspiration were indeed related through the phase shift of the closure expressed by changes in s_1. More specifically, when s_1 was negative (e.g. left panel in Figure 2.4), the closure in the HF signal was in an anti-phase relationship with the opening of the glottis, thus leaving more time for pre-aspiration (large A_{pre}) and less for post-aspiration (small A_{post}) to occur, and vice versa for positive s_1 (see e.g. the right panel in Figure 2.4 in which a closure is represented by low values in both HF and VP). The overall area $A_{tot}(s_1)$, i.e. the total amount of aspiration, remained stable across all values of s_1. Appendix A.3 provides proof that the steepness of the lines $A_{pre}(s_1)$ and $A_{post}(s_1)$ is high, while $A_{tot}(s_1)$ does not significantly depart from a flat line.

2.2.3 Discussion

The aim of this section was to identify whether there was a relation between pre- and post-aspiration strength, following the model in Figure 2.1. When FPCA

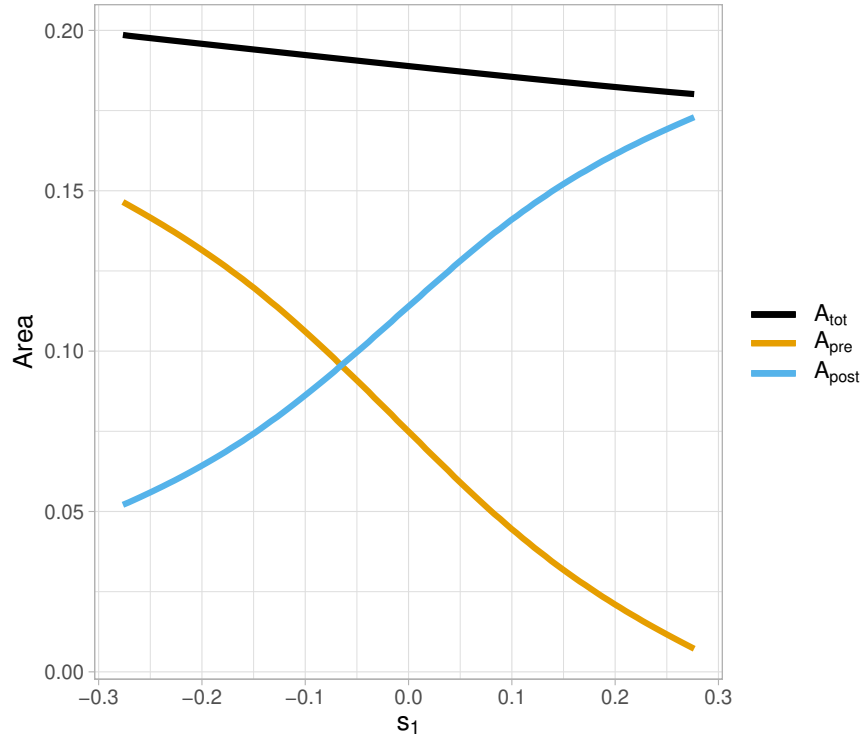


Figure 2.5: A_{pre} (yellow), A_{post} (blue), and $A_{tot} = A_{pre} + A_{post}$ (black) computed by using Eq. (2.2) as a function of s_1, when the signals HF and VP are defined as in Eq. (2.1) using only PC1.

was applied to HF and VP that had been extracted from the speech signals and that represent the oral constriction and glottal gesture respectively, it was found that the most important dimension for explaining the variance in these Andalusian data was the relative alignment of the closure which was captured by PC1 and modulated by score s_1. Moreover, a reconstruction of HF and VP using only PC1 in Eq. (2.1) (thereby eliminating all other sources of variation) showed a clear relationship between closure timing, pre-, and post-aspiration. Thus, later closures were associated with more extensive pre-aspiration and less extensive post-aspiration while for earlier closures pre-aspiration diminished and post-aspiration increased.

The results of this first part therefore provide some support for the model in Figure 2.1 based on a gestural model of speech production in which the extent of pre- and post-aspiration are inversely related to each other as a

consequence of how the closure is timed relative to the glottal opening. A major difference between the acoustic model presented here and the schematic outline based on articulatory gestures is that the former but not the latter takes amplitude into account. Thus, the model in Figure 2.1 is based only on timing considerations and not on the size of vocal tract opening during the intervals when the vocal tract is predominantly given over to aspiration. On the other hand, in the present study the areas between the acoustic signals of high frequency energy and voicing probability that were taken as a proxy for the extent of aspiration were influenced not only by the phasing of the closure, but also by the amplitude of energy in the HF signal. That is, Figure 2.4 has shown that PC1 did not only capture a phase shift of HF relative to VP, but also that post-aspiration had a higher energy peak than pre-aspiration. This functional analysis thereby presents more information on aspiration than would be possible using duration measurements based on a segmentation of the speech signals. However, in order for the amplitude (and hence areas) to be comparable across tokens, global amplitude differences which were most likely caused by speaker-dependent variations in the amplitude of the signal (caused e.g. by speaking more quietly or softly and/or as a result of different distances from the microphone) had to be factored out (see Appendices A.4 and A.5 for further details).

2.3 Influence of Speakers' Age and Region on Closure Phasing

The analysis suggesting that pre- and post-aspiration are predictable consequences of resynchronising the closure with the voiceless interval has been based so far on a model applied to the entire data set across all speakers and repetitions. The concern of this section is to test whether s_1, which was shown to capture variations in closure timing (section 2.2), also distinguishes between age and regional origin of the speakers. The prediction is that it should do so, given that older and East Andalusian speakers have been shown to have greater pre-aspiration and less post-aspiration than their younger and West

Andalusian counterparts (Ruch & Harrington, 2014; Ruch & Peters, 2016). If the prediction holds, then this would establish a link between synchronic variation and a sound change in progress.

2.3.1 Method

A linear mixed effect regression model was constructed for the first PC score s_1 as response variable and age (two levels), region (two levels), and cluster type (three levels) as fixed factors, while word (52 levels) and speaker (48 levels) were added as random factors. All statistical results reported below for s_1 also approximately hold for A_{post} (and for A_{pre} with negative sign), since the areas are near-linearly related to the first PC score (cf. Figure A.1 and see Appendix A.3 for further mathematical details). While the areas were helpful in demonstrating the existence of a relationship between pre- and post-aspiration in Andalusian Spanish, s_1 was chosen as response variable here because it directly expresses the closure phasing which, according to the model in Figure 2.1, is responsible for this relationship.

The full LMER model is given in Eq. (2.3) below (R notation):

$$s_1 \sim (age + region + cluster)^3 + (cluster|speaker) + (age + region|word) \quad (2.3)$$

where the term $(age + region + cluster)^3$ indicates the presence of three fixed factors plus all the possible two- and three-way interaction terms formed by them, while the random factor *speaker* is modulated by *cluster type* and *word* is modulated by *age* and *region*. The model was pruned using the R package `lmerTest` (version 3.1.2) in order to remove all non-significant factors and factor combinations. After pruning, all fixed and random terms were retained apart from the three-way interaction of the fixed factors and the two-way interaction between age and region. All post-hoc tests were computed using the R package `emmeans` (version 1.4.6).

In order to translate the predicted s_1 values back into HF and VP curves, the resulting estimated marginal means (EMMs) were substituted into Eq. (2.1),

where only s_1 was used and the other scores were set to zero. For instance, $HF(t)$ for /st/ produced by young West Andalusians was represented by $\mu_{HF} + s_1 \cdot PC1_{HF}(t)$, where s_1 took the EMM value for that particular factor combination.

2.3.2 Results

The boxplots in Figure 2.6 provide a graphical impression of how age, region, and cluster type affect the first PC score. Compatibly with the production of an earlier closure accompanied by post-aspiration, s_1 was higher in Figure 2.6 for younger speakers in all conditions. Figure 2.6 suggests only few region differences except for alveolars. The results of the mixed model with the fixed and random factors given in section 2.3.1 showed a significant main influence on s_1 of age ($F[1, 46.9] = 24.6$, $p < 0.001$) but not of region nor of cluster type. There was however a significant two-way interaction between age and cluster ($F[2, 54.4] = 6.6$, $p < 0.01$) as well as region and cluster ($F[2, 53.3] = 5.7$, $p < 0.01$). Post-hoc alpha-adjusted Tukey tests showed that there were significant differences between older and younger speakers for all three places of articulation (/sp/: $t = 3.8$, $p < 0.001$; /st/: $t = 5.0$, $p < 0.001$; /sk/: $t = 4.6$, $p < 0.001$). The post-hoc tests revealed no significant region differences. However, there was a trend towards a significant difference between speakers from East and West Andalusia producing the alveolar cluster ($t = 1.9$, $p = 0.06$). For young West Andalusian speakers there was a significant difference between labial and alveolar clusters ($t = 3.9$, $p < 0.001$).

Figure 2.6 also shows the estimated marginal means (EMMs) for each combination of age group × region × cluster type (black dots within the boxes), with their respective confidence intervals (vertical bands), which are the values of s_1 that the model predicts for each combination of the fixed factors. As described in section 2.3.1, these EMM values for specific factor combinations were used to reconstruct the HF and VP curves. This reconstruction (Figure 2.7) shows that the closure (represented by the HF signal) occurred earlier for younger (green) than for older (dark grey) speakers for all three places of articulation and both regions. These differences between the age groups are the consequence of (i) $s_1(\text{young}) > s_1(\text{old})$ as found by the LMER model (cf.

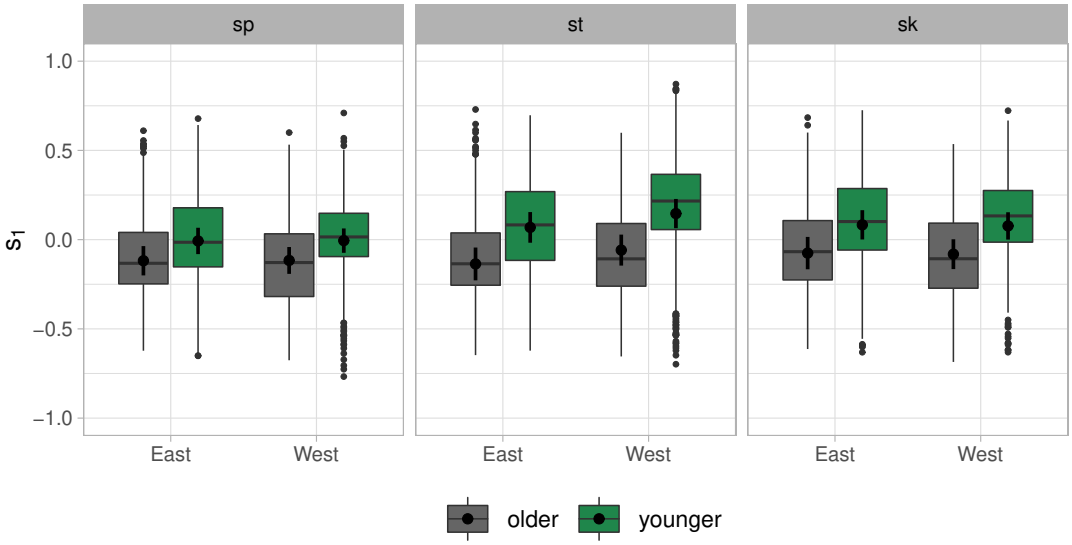


Figure 2.6: Boxplots of s_1 values as well as estimated marginal means of s_1 (black dots within the boxes) with related confidence intervals (black vertical bands around the dots) based on Eq. (2.3). Younger speakers are shown in green, older ones in dark grey.

Figure 2.6) and (ii) s_1 mainly modulating a phase shift of HF, where the shift is towards the left of the time axis for increasing s_1 (cf. Figure 2.4). Additionally, the HF signals for /st/ and young speakers differed from each other between East and West, the latter being more left shifted and reaching higher energy values in the second part of the signal. Although this difference was not found to be significant, it is clearly visible in the reconstructed curves.

2.3.3 Discussion

There were age- and (to a lesser extent) region-dependent variations in the PC score s_1 derived from FPCA. These differences between the speaker groups together with the evidence that PC1 models a phasing of the closure suggest that, consistently with various other studies (Moya Corral, 2007; O'Neill, 2010; Ruch, 2013; Torreira, 2006, 2007), there is a sound change in progress in /sC/ clusters in Andalusian Spanish. Based on the analysis and conclusions in section 2.2, the direction of change is such that the closure is timed to occur earlier in younger than in older speakers as a consequence of which

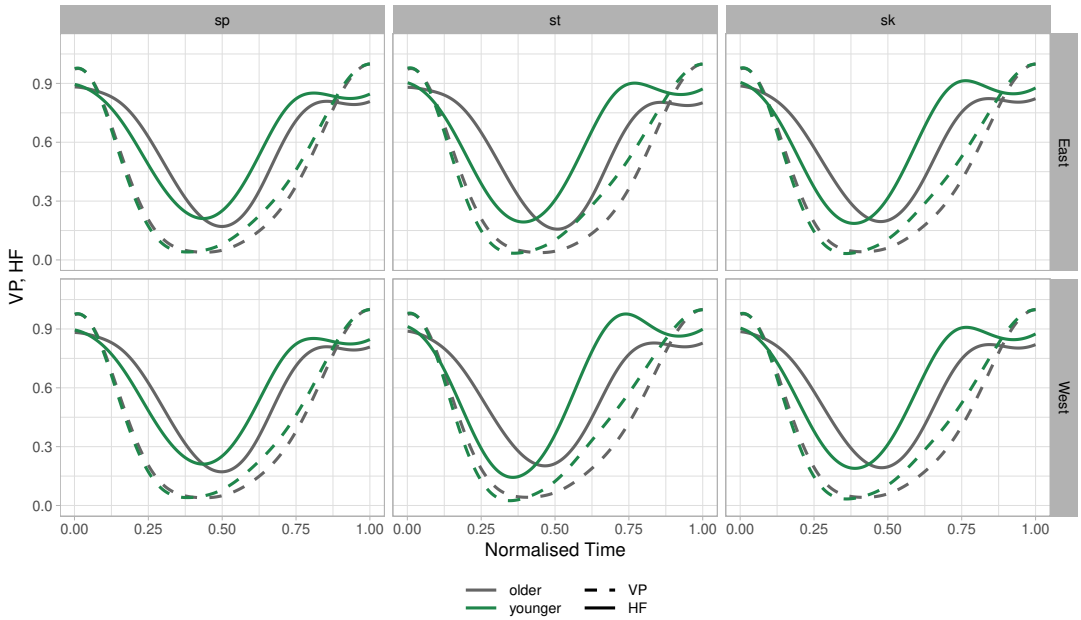


Figure 2.7: Reconstruction of $HF(t)$ (solid) and $VP(t)$ (dashed) using Eq. (2.1). For each factor combination, EMMs are used for s_1, while $s_k = 0$ for $k > 1$. The curves for the younger age group are green, the ones for the older group are dark grey.

younger speakers produce these clusters with more post-aspiration and less pre-aspiration than their older counterparts.

The East Andalusian variety is more conservative as far as this sound change is concerned than its Western counterpart, as others (O'Neill, 2010) have shown. A comparison between the two regions can therefore provide some clues about how sound change is affected by phonetic context. The analysis conducted in this section suggests that alveolar contexts might lead the sound change while labial clusters seem to be the last ones to be affected by the sound change. This is because, firstly, there was a trend towards a significant difference between East and West Andalusian speakers only for /st/ and secondly, there was a significant difference between /sp/ and /st/ for the most advanced speaker group, namely young West Andalusian speakers. This finding is compatible with the suggestion in Ruch and Peters (2016) that the sound change by which pre- evolves into post-aspiration first originates in the alveolar context before spreading to the velar and, lastly, to the labial context.

2.4 General Discussion

The first part of this study showed that pre- and post-aspiration strength in Andalusian Spanish are inversely related to each other and a predictable consequence of how the closure is phased with respect to the voiceless interval in V_1sCV_2 sequences. This finding is consistent with the model based on articulatory phonology proposed by Parrell (2012) in which closure rephasing is at the core of the change from pre- to post-aspiration at a faster speech rate in the production of /st/ in *pastándola*. It also extends this model by suggesting that the inverse relationship between pre- and post-aspiration as a consequence of closure re-phasing depends not just on timing but also on scaling, i.e. on the relationship between the amplitude of aspiration noise before and after the closure.

The second part of the study showed that the closure phasing and hence the inverse relationship between pre- and post-aspiration are conditioned by age and to a lesser extent by region. As far as age is concerned, these results are compatible with evidence showing a sound change in progress in Andalusian Spanish (O'Neill, 2010; Ruch & Harrington, 2014; Ruch & Peters, 2016) given that the closure was found to be timed earlier for younger than older speakers. The change in progress was more advanced for speakers from West than from East Andalusia when producing the alveolar cluster. The new approach in this study is that these age, region, and place of articulation differences have been established based on analyses of pairs of time-varying signals requiring no segmentation of the closure and aspiration intervals. This approach has a methodological advantage given that these intervals (in particular V_1 and pre-aspiration) overlap with each other, thereby often making vertical segmentation unreliable and inconsistent (Fowler & Smith, 1986). Speech is inherently dynamic involving the overlapping of autonomous articulatory gestures (Fowler & Saltzman, 1993) and regular sound change almost always arises out of dynamic processes such as coarticulation (J. J. Ohala, 1993a) and undershoot (Lindblom et al., 1995). The method proposed in this study based on FPCA is appropriate for modelling these dynamic aspects of speech and the sound changes that they give rise to precisely because

it provides a way of categorising speech signals and of quantifying change without having to enforce an often arbitrary vertical segmentation of the speech signals (Fowler, 1984; Fowler & Smith, 1986). Although FPCA does not produce a dynamic model in the strict sense (e.g. a set of differential equations), it provides a way of isolating the variation in the pair of signals $HF(t)$ and $VP(t)$ (as shown by PC1) corresponding to the gestural dynamics sketched in Figure 2.1. In other words, the evidence that the main variation in those signals is due to a phase shift emerges from the statistical FPCA of the signal shapes.

Figure 2.8 is a summary of the dynamic approach to analysing the Andalusian database and its potential association to components of a cognitive-computational model of sound change (Ettlinger, 2007; Stevens et al., 2019; Todd et al., 2019; Wedel, 2006). In this study, FPCA (section 2.2.1.3) was applied to a database of dynamic episodes of speech derived from acoustic speech signals in order to obtain a signal decomposition model that consists of PC curves and PC scores. There are two different kinds of information that follow from the decomposition: First, the model shows how the time-varying signals that represent articulatory gestures systematically vary in shape and phasing. Second, the decomposition assigns every dynamic speech episode i its specific PC scores $s_{1,i}, s_{2,i}, \ldots, s_{k,i}$. All PC scores of all tokens in the database form a distribution in an abstract, multidimensional space. The location of the scores in this space can depend on speaker-specific properties like their age or regional origin, as shown in the second part of the study (section 2.3).

The cognitive-computational architecture of speech processing that is proposed in this model adopts the idea from exemplar models that phonological categories stand in a stochastic relationship to remembered speech signals (K. Johnson, 1997, 2006; Pierrehumbert, 2003a, 2006). The central idea here is that in human speech processing, individuals derive both phonological and distributional information after applying a transformation such as FPCA to multidimensional, remembered, time-varying episodes of speech. The derived phonological knowledge is analogous to separate tiers in articulatory phonology containing independent gestural dynamics showing (as in Figure 2.4) how the shape and phasing of gestures that characterise the phonological

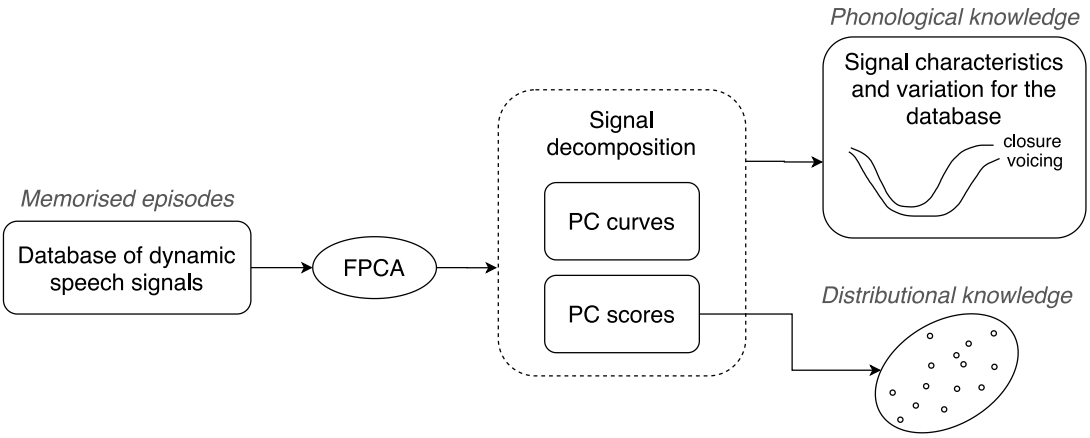


Figure 2.8: A schematic outline of the outputs after applying FPCA to the database in this study and their potential association (in italics) to components of a cognitive-computational model of sound change. The PC-based signal decomposition model gives rise to both summary signal characteristics over the entire database as well as distributional information.

category vary across the population of speakers. The distributional information is a cloud of points derived from the remembered episodes for the same phonological category (cf. Figure 2.6).

The transformation of memorised episodes of speech (e.g. FPCA in Figure 2.8) can result in a large amount of dimensions along which signals of the same phonological category can vary (e.g. PCs). The further issue to be considered is how individuals learn which kinds of variation are most relevant for a given phonological category. We suggest this might be guided by both phonetic and phonological criteria. The *phonetic* criterion is that a dimension of variation represents how the tokens of the phonological category are actually produced. In the case of the Andalusian database presented here, the most relevant variation for the /sC/ category was how the opening/closure phase of the vocal tract was variably timed with respect to a voiceless interval (i.e. PC1). The *phonological* criterion for identifying a dimension of variation as relevant is that it represents a group-level characteristic of the phonological category. That is, for the Andalusian data, the level of aspiration in episodes of /sC/ must be high regardless of how the closure is phased relative to the

voiceless interval. This is evident in the black line in Figure 2.5 which shows that the total amount of aspiration (expressed by A_{tot}) is more or less constant.

Such knowledge is functionally useful because it is likely to be a feature that distinguishes aspirated from non-aspirated clusters. Although we have not investigated unaspirated stops in the present study due to sparse data for /p, t, k/, our earlier investigations (Ruch & Harrington, 2014; Ruch & Peters, 2016) showed that the extent of pre- and post-aspiration is much less in unaspirated than aspirated clusters. Thus, whereas we find high levels of aspiration irrespective of how the closure is timed in the aspirated cluster of e.g. *pasta* (engl. pasta), the corresponding black line in Figure 2.5 is likely to be much lower for the unaspirated /t/ in *pata* (engl. paw). It is from this perspective that it is of functional value to choose a dimension of variation that is likely to provide categorical information for distinguishing between aspirated, i.e. /st/, and unaspirated, i.e. /t/, plosives. The actual classification i.e. distinction between an aspirated and unaspirated cluster would be accomplished not directly by this phonological information but instead by calculating the probability of class membership to the cloud of data points that are also derived by FPCA. Thus the model in Figure 2.8 shows, compatibly with exemplar models, that there is a stochastic relationship between phonological knowledge and speech signals: the new angle proposed here is that a transformation analogous to FPCA is intermediary between the two.

A sound change in progress can sometimes be characterised by a perceptual trading relationship between the coarticulatory source and effect. Beddor and colleagues (Beddor, 2009, 2012; Beddor et al., 2018) have investigated the phonologisation of nasalisation in the vowels of words exemplified by American English *send* and *sent* from this perspective. They show that there is an inverse relationship in speech production between the extent of vowel nasalisation and the duration of the following /n/ that gives rise to coarticulatory nasalisation. In perceiving *send*, they also show that listeners often identified nasalisation from the signal without associating or parsing it explicitly with either the vowel or following nasal consonant. This trading relationship is an appropriate strategy in perception for such variation, given that nasalisation in the American English variety that they investigated could be manifested in

the vowel, in the following nasal consonant, or both in speech production (and variably so between listeners in speech perception).

There is a degree of commonality between Beddor's findings and those in the present speech production study of Andalusian Spanish in which there was shown to be an inverse relationship between pre- and post-aspiration (Figure 2.5). This inverse relationship in our study only emerges, however, following the application of an FPCA transformation across speech signals from several speakers that differed in the extent to which they produced /sC/ clusters with pre- or post-aspiration. Without this FPCA transformation, there is no such inverse relationship. This is demonstrated by Figure 2.9 in which the areas A_{pre} and A_{post} were calculated from Eq. (2.2) where HF and VP were directly obtained from the speech signals, without applying FPCA. The general conclusion from Figure 2.9 is that the inverse relationship between pre- and post-aspiration is not directly manifested in the acoustics of any (or several) /sC/ clusters. It is perhaps for this reason that there has been scant evidence for such a relationship from other studies (Ruch, 2013; Ruch & Harrington, 2014; Ruch & Peters, 2016; Torreira, 2007). The inverse relationship exists instead at a more abstract level that is a consequence of modelling the acoustic speech of multidimensional, time-varying speech signals over many words and repetitions and above all over many speakers of which some have predominantly pre- and others predominantly post-aspiration in producing these clusters (see Appendices A.4 and A.7 for more details).

An individual who has abstracted the phonological and distributional knowledge from the type of data investigated in this study is likely to classify pre- or post-aspiration equivalently. This is apparent in Figure 2.5 which shows that, irrespective of whether the closure is late or early, the quantity of aspiration stays more or less the same. Here there are parallels once again to Beddor's findings showing that, at least for some listeners, it does not matter whether the nasalisation occurs in the vowel or the following consonant: both are treated equivalently as being [+nasal]. Whether or not Andalusian listeners actually exhibit such trading relationships is not something that we have yet investigated. Following analogous nasalisation studies by Beddor (2009) and Zellou (2017), we would expect a considerable degree of variation

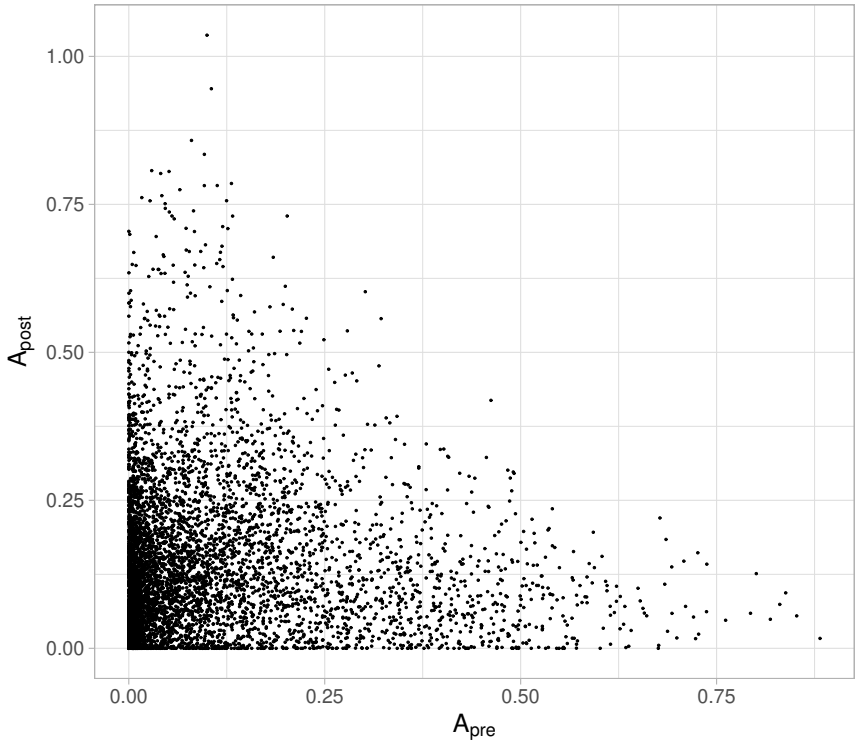


Figure 2.9: Area A_{post} against A_{pre} as in Eq. (2.2) when $HF(t)$ and $VP(t)$ are the (smoothed and time-normalised) curves obtained directly from the speech signal without FPCA transformation.

across Andalusian listeners in whether or not such a trading relationship is manifested. The further prediction from the type of model in Figure 2.8 is that experience conditions whether or not a listener demonstrates such a trading relationship. Listeners who have been exposed predominantly to old, Eastern Andalusian speakers (who typically pre-aspirate) or those exposed mostly to younger, Western Andalusian speakers (who typically post-aspirate) are predicted to show much less evidence of a trading relationship than those listeners exposed to both these types of speakers.

The model in Figure 2.8 brings together insights from articulatory phonology and episodic models of speech in order to relate synchronic variation to diachronic change. Articulatory phonology has provided great advances in understanding speech dynamics, but given its historical emphasis on articulatory invariants (Fowler, 2003), has been not so easily adaptable to the findings in

the last 20 years or so that the relationship between phonological knowledge and speech is a stochastic one. By contrast, exemplar theory has provided great advances in modelling this stochastic relationship. With few exceptions (e.g. Kirchner et al., 2010), there has, however, been an almost complete neglect in explaining quite how these stochastic phonological categories are derived and associated with multidimensional speech signals that change in time. The idea in the present study that users of the language may extract dimensions across their remembered speech signals brings together these important insights from these separate models. This unified model shares with episodic models of speech that there is no sharp distinction either between synchronic variation and the resulting diachronic change nor between phonological knowledge and the (remembered) speech signals out of which such knowledge is constructed.

3 | An Agent-Based Model of Sound Change: soundChangeR

Abstract

This chapter is concerned with a cognitively-inspired agent-based model (ABM) of sound change, i.e. a computational model in which human speakers are represented by computational agents. These agents are given rules and mechanisms in order to produce and perceive acoustic exemplars of words as well as a memory in order to store them. Exemplars are associated to a word class, and exemplars and words are flexibly linked through a phonological class. In the latest version of the ABM, the phonological classes are agent-specific as they are regularly recomputed using two unsupervised machine-learning algorithms. Using simulations of artificially generated data, I aim to show how sound change can arise as a consequence of the interplay between the input data, the implemented mechanisms, and the prolonged interactions between agents, as predicted by the interactive-phonetic model of sound change. The final part of this chapter is about the ABM's characteristics and their implications also in comparison to other computational models of sound change as well as about the relation between simulated and quantitative results.

This chapter is partially based on and complements Gubian et al. (2023). See Appendix B.1 for the authorship contribution statement for the published article.

3.1 Introduction

This chapter is concerned with the architecture and core properties of a cognitively-inspired agent-based model (ABM) of sound change. In ABMs, agents interact with one another along a defined set of rules and it can be observed how individual actions (micromotives) lead to population-wide patterns (macrobehaviours; Schelling, 1978). That is, the power of these models is to demonstrate the emergence of global phenomena out of local decisions which makes them especially useful for the investigation of complex adaptive systems (Bankes, 2002; Beckner et al., 2009; Berry et al., 2002; Bonabeau, 2002). A complex system consists of many individual parts which are interdependent "so that the emergent behavior of the whole is difficult to predict from the behavior of the parts" (MacLennan, 2007, p. 173). When the parts of a complex system can change as a response to their environment, the system is adaptive. According to Beckner et al. (2009), spoken language qualifies as a complex adaptive system because it fulfils four key characteristics:

> (a) The system consists of multiple agents (the speakers in the speech community) interacting with one another.
>
> (b) The system is adaptive; that is, speakers' behavior is based on their past interactions, and current and past interactions together feed forward into future behavior.
>
> (c) A speaker's behavior is the consequence of competing factors ranging from perceptual mechanics to social motivations.
>
> (d) The structures of language emerge from interrelated patterns of experience, social interaction, and cognitive processes.
>
> (Beckner et al., 2009, p. 2, line breaks added)

These four features of language as a complex adaptive system are all integral concepts implemented in the agent-based computational model presented in this chapter, as will be shown in section 3.2. Importantly, feature (d) implies that language change – and, by extension, sound change – can be modelled as a consequence of the interplay between the other three features, which is what we attempt to do with the ABM.

The motivation for developing such a model for the study of sound change is twofold: First, an ABM provides a controlled environment which can be used to test the factors that might play a role in the emergence and spread of a specific change that has already been observed empirically. This is necessary because it is impossible to know beforehand whether a sound change is going to take place, so as soon as the change is underway, it is too late to capture the circumstances which may have triggered it. In the artificial world of the agent-based model, however, it is possible to explore how both intra- and extralinguistic factors may have contributed to the occurrence and progression of a sound change. The second motivation is to test and further develop theories of sound change. Every ABM relies on assumptions which are drawn from evidence-based theories of sound change. If such an ABM is capable of successfully replicating different kinds of sound changes, it provides strong support for the theoretical model. If, on the other hand, the results of a simulation fail to replicate a sound change, a careful analysis of what has gone wrong may provide new insights and lead to an adaptation of the tested theory of change. Once an ABM has been shown to successfully replicate several unrelated sound changes, it can even help to inform a holistic model of sound change. ABMs are hence a computational method of shedding "light on the architecture of human speech processing, how it is flexibly adapted to social variation in language, and which mechanisms within this architecture can give rise to change" (Harrington et al., 2018, p. 3).

The ABM presented here is a computational implementation of the interactive-phonetic (IP) model of sound change (Harrington et al., 2018). The IP model tries to unify the two major strands of sound change research that, so far, have mostly been pursued independently of one another (Harrington & Schiel, 2017; Stevens & Harrington, 2014): the first is concerned with the origins of sound change (e.g. Lindblom et al., 1995; J. J. Ohala, 1989, 2012), while the second asks how a new variant of a sound can spread among the members of a speech community (e.g. Eckert, 1988; Labov, 2001; Trudgill, 2004). Connecting these two approaches to the study of sound change may be the key to answering one of the most pressing questions in this field which remains largely unsolved despite an upsurge of research (Baker et al., 2011;

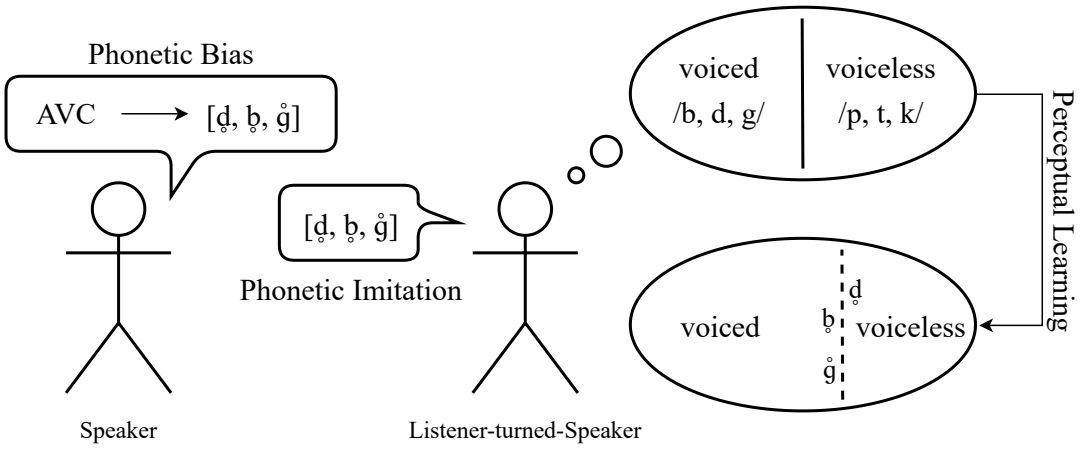


Figure 3.1: Schematic sketch of the three focal points of the IP model: phonetic bias (as exemplified by the aerodynamic voicing constraint (AVC)), perceptual learning, and phonetic imitation.

Bermúdez-Otero, 2020; Kirby & Sonderegger, 2015; Sóskuthy, 2015; Stevens & Harrington, 2014; Yu, 2013): Why does sound change happen under one set of circumstances, but not another? More precisely, why does a sound change occur in one language but not another, or in the same language but at a different point in time? This so called actuation problem as posed by Weinreich et al. (1968) requires an investigation of the complex mixture of intra- and extralinguistic factors which affects the emergence and progression of sound changes. In order to link the models concerned with the origin and spread of sound changes, the IP model focuses on four related concepts, the first three of which are also pictured in Figure 3.1: phonetic biases, non-social phonetic imitation in stochastic interactions between individuals, perceptual learning, and the flexible association between word classes and memorised traces of speech.

Phonetic biases are the result of the processes of speech production and speech perception (see Garrett and Johnson, 2013 for an overview). That is, the variability introduced by these processes is often not random, but directional and asymmetric. In Figure 3.1, for instance, a speaker has produced the originally voiced plosives /b, d, g/ as their devoiced variants /b̥, d̥, g̊/ because of the aerodynamic voicing constraint (AVC). The aerodynamic constraint on voiced plosives is a well-known example of directional synchronic variation

(J. J. Ohala, 1983, 1989). In order to produce a voiced sound, there must be a constant flow of air through the glottis. In the case of a voiced plosive the air is collected in the oral cavity behind the place of the closure, thereby increasing the supraglottal air pressure. If the speaker does not release the closure for some time, the air pressure above and below the glottis become so similar that the air flow, and hence the vibration of the vocal folds, is stopped, resulting in a voiceless plosive. That is, voicing cannot be maintained through a long closure which creates a bias towards voiceless plosives (J. J. Ohala, 1997). This constraint along with many other biases create a "pool of synchronic variation" (J. J. Ohala, 1989) which can provide the raw material for sound change (although only a fraction of such variation is eventually turned into sound change).

Interactions between individuals play a seminal role in triggering and propagating a sound change (Labov, 1963; Trudgill, 1999, 2008a). According to the IP model, this is because humans tend to imitate a conversational partner's (linguistic) behaviour which can turn a stable phonetic bias into unstable change (Harrington et al., 2018). Phonetic imitation is defined as an individual's subconscious adjustment of acoustic speech characteristics towards those of an interlocutor (Delvaux & Soquet, 2007; Pardo, 2006, 2013; Sato et al., 2013). If the speaker in Figure 3.1 consistently devoiced /b, d, g/, the listener-turned-speaker might imitate this phonetic behaviour (Y. Lee et al., 2021; Nielsen, 2011; Zellou & Brotherton, 2021) – even though this process takes place below the threshold of consciousness and the listener-turned-speaker would not be aware of any changes in their speech production (Lakin & Chartrand, 2003; Pardo et al., 2012). That is, phonetic imitation is a subconscious and imperceptible process (Garnier et al., 2013; Kappes et al., 2009) that occurs regardless of social factors (Shockley et al., 2004), although e.g. attraction can affect the degree of convergence (Abrego-Collier et al., 2011; Babel, 2012; Pardo et al., 2012). However, phonetic imitation is different from social accommodation (Giles, 1973) which is explicitly not taken into account by the IP model. It is rather a more general pattern of human behaviour, given that it also occurs in other domains such as synchronisation in clapping (Néda et al., 2000), limb movement (Richardson et al., 2005), and even brain rhythm

(Kawasaki et al., 2013). It has been hypothesised that phonetic imitation – in contrast to social factors such as the need for a common identity – is one of the essential mechanisms in the emergence of new dialects (Sebanz et al., 2006; Trudgill et al., 2000) and can permanently alter phonological systems (Nguyen & Delvaux, 2015). Phonetic imitation also supports theories which propose a strong cognitive link between speech production and speech perception.

So if, in a conversation, a speaker tends to devoice their voiced plosives because of the aerodynamic constraint, this may be picked up and imitated by the interlocutor (usually for the duration of the conversation or until shortly thereafter; Eisner and McQueen, 2006). If the interlocutor encounters many devoiced plosives over time, they might incorporate this supposed feature of the language into their own speech repertoire. The process which makes such a shift within an individual listener-turned-speaker possible is called perceptual learning (Clarke-Davidson et al., 2008; Norris et al., 2003). That is, ambiguous speech signals can lead to a shift of perceptual category boundaries which impacts the phonological classification of perceived speech signals (Saltzman & Myers, 2021; Samuel & Kraljic, 2009). In contrast to phonetic imitation, perceptual learning involves intra-individual changes in how the mental lexicon in structured phonologically. In the example shown in Figure 3.1, the individual on the right initially differentiates between /b/ and /p/, /d/ and /t/, and /g/ and /k/ on a phonological level. Because the individual perceived many ambiguous tokens of /b, d, g/ – namely tokens that were devoiced – the perceptual boundary between voiced and voiceless plosives shifts towards the voiceless category. In order for perceptual learning to occur, the individual must perceive the same ambiguity in many tokens from many speakers (and not just the three exemplary tokens from one speaker displayed in Figure 3.1), i.e. whether or not perceptual boundaries shift permanently depends on the stochastic contact between a large group of individuals, some of which consistently produce ambiguous tokens e.g. because of a phonetic bias like the AVC.

Both non-social phonetic imitation and perceptual learning are predicted by episodic models of speech (Goldinger, 1996, 1998; Palmeri et al., 1993) and exemplar theory (e.g. Pierrehumbert, 2001) which state that language

experience shapes both speech production and speech perception. More specifically, these models state that perceived speech signals are parameterised and stored mentally in a multi-dimensional phonetic space. Also in line with exemplar theory, the IP model proposes that the association between word classes and remembered speech signals through phonological classes is probabilistic, flexible, and can differ from individual to individual (Harrington et al., 2018). In Figure 3.1, for instance, the listener-turned-speaker has developed phonological classes that are different from the canonical separation between /b, d, g/ on the one hand and /p, t, k/ on the other and will keep changing the association between word classes and remembered speech signals through phonological classes with growing language experience. The IP model also supports the notion of sub-phonemic classes, i.e. classes which hold allophonic rather than phonological information. This is firstly because these classes are created bottom-up as an abstraction over a cloud of parameterised and memorised speech signals which are specific to each individual (Pierrehumbert, 2001, 2003a, 2003b), and secondly because it has been shown that sub-phonemic classes are of relevance in speech perception (German et al., 2013; Jones & Clopper, 2019; Luthra et al., 2019; Mitterer et al., 2013; Nielsen, 2011; Reinisch et al., 2020; Reinisch & Mitterer, 2016; Scobbie & Stuart-Smith, 2008). In terms of the example in Figure 3.1, the listener-turned-speaker may over time create sub-phonemic categories for voiced, voiceless, and devoiced plosives (where the latter are not distinctive, and hence not phonological but sub-phonemic). So in summary, according to the IP model which provides the theoretical basis for the agent-based model presented here, "whether or not sound change actually comes about depends upon which speakers regularly speak to each other, whether a phonetic bias happens to be magnified by interaction, and whether or not sub-phonemic classes are fragmented and regrouped over time" (Harrington et al., 2018, p. 17; also see Dediu and Moisik, 2019).

The rest of this chapter is structured as follows: Section 3.2 is concerned with explaining all concepts, entities, and processes of the agent-based model which was implemented as an R package called `soundChangeR`. In section 3.3, the core mechanisms of the ABM and their implications are demonstrated using simulations of artificially generated data. The strengths and weaknesses

of the model are discussed and compared to other computational models of sound change in section 3.4.

3.2 Computational Implementation

The first computational implementation of the IP model was presented by Harrington and Schiel (2017) where the shift of /u/ to the front of the vowel space was simulated based on data from speakers of Standard Southern British English (SSBE). Since then, this agent-based model has been shown to successfully model other sound changes, but it has also evolved over time. We refrain from describing the evolution of the model and instead focus on the latest version which has been published as an R package of the name `soundChangeR` on GitHub (https://github.com/IPS-LMU/soundChangeR). At the time of writing, the current version of `soundChangeR` is 1.0.0. Basic installation instructions for the R package are given in Appendix B.2, but the reader is referred to the GitHub repository for more details. The vignette to `soundChangeR` in Appendix B.3 contains detailed explanations of and all relevant information about the software side of this ABM.

This section is instead about concepts, constraints, and entities that play a role in this model, the central one being the agents. Agents are the computational instantiation of human speakers and listeners. That means that the agents are equipped with mechanisms in order to produce and perceive speech as well as a memory that connects the two processes. At the beginning of a simulation, each agent is given real production data to initialise their memories. The memory is a storage for exemplars which consist of the parameterised acoustics of a sound as well as the word in which the sound was uttered (e.g. formant values for the vowel in the word *food*). In between the phonetic-acoustic space and the lexical level of the memory, there is a phonological level which can either be pre-determined by the user (e.g. /u/ for the vowel in *food* and /i/ for the vowel in *feed*) or calculated by means of two machine learning algorithms. When pre-determined, the phonological classes remain fixed over the course of the simulation, thereby preventing changes in the association between exemplars and word classes via phonological classes. The

machine learning algorithms, on the other hand, recalculate the phonological classes regularly and separately for each agent. To do so, these unsupervised algorithms do not need any prior information about the exemplars apart from their location in the phonetic-acoustic space and their association to word classes. Given these premises – a memory filled with exemplars which are associated to lexical and phonological classes – the population of agents starts to interact, i.e. an agent speaker produces an exemplar and an agent listener perceives it. The production of exemplars is a Gaussian sampling procedure, i.e. a new exemplar is sampled from a Gaussian distribution which was computed over all exemplars belonging to a chosen word class. The perception (or rather memorisation, given that word recognition and lexical access are not modelled) is much more constrained and follows two criteria in order to decide whether the perceived exemplar should be memorised: The first criterion tests the exemplar's typicality, i.e. whether it is located close enough to the intended phonological class in the agent listener's acoustic space. The second criterion tests whether the exemplar is discriminable, i.e. whether it has a higher probability of belonging to the intended than to the competing phonological classes. So while production is a word-based process, the decision of memorising an exemplar in perception is based on phonemes, and both processes draw information from the same memorised pool of exemplars. In order to limit the amount of exemplars stored in the memory, agents can forget exemplars by removing them from memory. So in summary, agents initialised with human production data can exchange exemplars in a production-perception feedback loop which, over time, can result in acoustic and phonological changes. All concepts briefly mentioned here are explained in detail below (for all implicit assumptions and simplifications of the model, see Appendix B.7).

3.2.1 Agents and Exemplars

An agent is the computational representation of a real person, i.e. the agent is initialised with the speech characteristics of an actual speaker (but see section 3.2.2). Since this model aligns more with the mechanistic view on sound change (Trudgill et al., 2000) and focuses less on the sociolinguistic

components of such changes (Labov, 2001), the only social attribute of an agent is their agent group. This grouping of agents can depend on the speakers' age, regional origin or any other sociolinguistic variable that may be relevant to the observed change. Dividing the agent population into groups is not obligatory, and if the user decides to define agent groups, they do not have to be binary. The agent groups impact the choice of an agent speaker and an agent listener during the interactions: agents can interact only within their own group, across groups, or they can interact randomly with one another regardless of their group. An interaction always consists of an agent speaker who produces a new token of a word and an agent listener who decides whether or not to memorise the token.

Agents are given a memory to store traces of speech which are then used for both speech production and perception. In line with Exemplar Theory, we call these traces exemplars (or, alternatively, tokens). Exemplars are a usually vectorial acoustic representation of the speech sounds under investigation. The only technical requirement for the acoustic features is that they must be numeric and continuous, e.g. they can be formant values, durations, or parametric representations such as DCT coefficients and principal component (PC) scores. The number of acoustic features used as dimensions of the phonetic space in the agents' memories is up to the user. Exemplars also have a fixed association to the word in which they were uttered, so word classes are statistical generalisations over clouds of acoustic exemplars. The phonological level in the ABM provides a link between the exemplars and word classes as explained in section 3.2.5.

3.2.2 Initialisation of Agents

Usually, every human speaker is represented by one agent which is achieved by initialising the agent with the acoustic data of that speaker. However, it is possible to break this paradigm by applying bootstrapping. In the context of the agent-based model, bootstrapping means that the population is created from the pool of real speakers, but one speaker can be represented by several agents. If, for example, 20 speakers were recorded but the agent population

should consist of 100 agents, the 20 speakers are randomly sampled until 100 agents are initialised. Since this is a random process, the speakers are not necessarily represented equally in the agent population (e.g. of the 20 speakers, some may have informed more agents than others). In order to avoid exact clones of the speakers, the ABM provides options to create new exemplars by means of a resampling technique (see section 3.2.6) and remove the agents' original exemplars. It is also recommended to complete multiple runs of a simulation with a bootstrapped agent population such that the settings are the same, but the bootstrapped population is different every time. This is necessary in order to avoid spurious results and to test for the robustness of the simulation's outcome. Bootstrapping can also be used to manipulate the amount of agents per agent group (e.g. more agents of group A than of group B).

3.2.3 Production

The production algorithm in `soundChangeR` is a relatively simple Gaussian sampling procedure which does not explicitly model the execution of motor plans or articulatory processes. Instead, the agent speaker randomly chooses a word class, builds a Gaussian distribution over all memorised exemplars associated with that word, and samples a new token from it. Contrary to many other computational models of sound change, the produced token is not subject to a bias that would push it towards or away from an articulatory target. In `soundChangeR`, production is a word-based process in order to allow for possible coarticulatory effects to be carried over into the acoustics of the produced token. That is, if there was a coarticulatory effect of the phonetic context on the observed sound in the original speech of the human speaker, it can be reproduced by the word-based sampling procedure in the agent-based model. A long-term effect of Gaussian sampling in production together with the constraints on memorisation described in section 3.2.4 is that the original acoustic distributions become narrower over time because new tokens are more likely to be close to the mean of the distribution than to be on its

tails. Appendix B.6 provides a more in-depth example of the production and perception of a taken during an interaction between two agents.

3.2.4 Perception

The agent listener receives the acoustic token together with its associated word class and makes use of phonological knowledge in order to decide whether or not to memorise the token. This decision is not about word recognition: Since misunderstandings are considered neither a catalyst nor an obstacle to sound change, it is assumed in this model that lexical access and word recognition work perfectly. Thus, the term perception is used very loosely here and only refers to the process of updating an agent's memory. The agent listener's task in perception is strongly linked to the phonemic level, i.e. the memorisation criteria are computed with reference to the phonemic classes (see section 3.2.5).

There are two main memorisation criteria which were variably used in earlier versions of this model. The first criterion determines whether the Mahalanobis distance between the token and the centroid of the corresponding phonological class is lower than a given threshold. If so, the token is memorised. Hence, this criterion is a test of the token's typicality (also see e.g. Todd et al., 2019). Since the Mahalanobis distance threshold is computed only with regard to the intended phonological class of the perceived exemplar, this memorisation criterion is called absolute (in contrast to the relative criterion, see below). A natural consequence of this criterion is that widely spread or skewed phonological classes are more likely to incorporate new tokens than compact ones. An example of this is given in Figure 3.2 where two agents, A (left) and B (right), have different representations of the same phoneme classes P1 (solid) and P2 (dashed). The ellipses enclose the same probability mass in A and B, i.e., their contours define locations at the same Mahalanobis distance to their respective centres (plus signs). In this artificial example it is assumed that both agents are asked to evaluate the same token *t1* (red) as a potential member of phoneme P1. If both agents apply the absolute criterion to do so, the token would only be memorised by agent A. This is because *t1* is enclosed by agent A's phonological class P1, i.e. the Mahalanobis distance between *t1*

and the centroid of P1 is below the threshold indicated by the ellipse contours. The opposite is true for agent B: Since B's phonological class P1 is very narrow, *t1* would have to be much closer to the centroid in order for it to be memorised according to the Mahalanobis distance. Section 3.3.2.1 is concerned with the effects of the absolute criterion when applied during interactions between agents A and B.

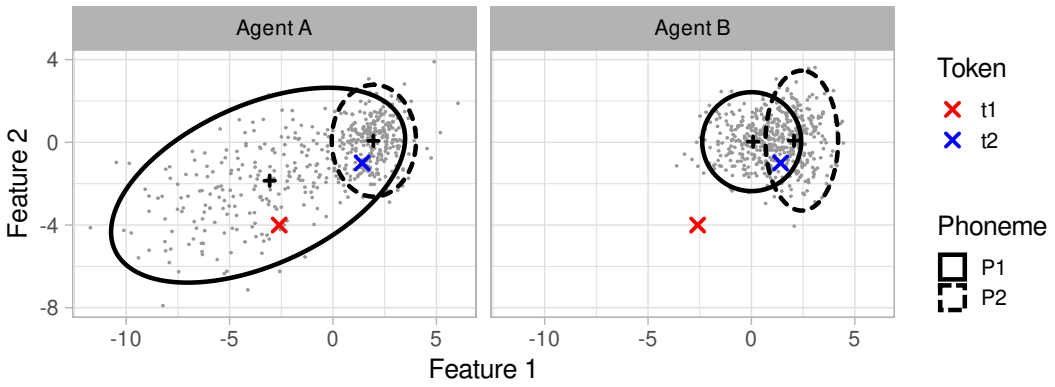


Figure 3.2: Artificially constructed example of phonological classes P1 (solid) and P2 (dashed) by two agents A and B in a two-dimensional acoustic space and their relation to tokens *t1* (red) and *t2* (blue). The grey dots are exemplars and the ellipses' respective centroids are marked by plus signs.

The second memorisation criterion is called relative because it takes all phonological classes in the agent listener's memory into account. In this case, the perceived token is only memorised if its posterior probability conditioned on the intended phonological class is higher than the posterior probabilities conditioned on the competing phonological classes.[6] Conceptually, using posterior probabilities in this way is a test of the token's discriminability (also see e.g. Todd et al., 2019). In the example given in Figure 3.2, agents A and B test whether token *t2* (blue) should be incorporated into the intended phoneme class P2 by applying the maximum posterior probability decision. For agent A, the posterior probability of *t2* given the phonological class P2 is 0.94, i.e.

[6] A variation of this so called maximum *a posteriori* probability criterion tests whether the posterior probability of the token conditioned on the corresponding phonological class exceeds a given threshold (e.g. Harrington & Schiel, 2017).

t2 is probabilistically much closer to P2 than to P1 and would therefore be memorised as an exemplar of P2. For agent B, on the other hand, *t2*'s posterior probability is 0.49 when tested for P2 (and, hence, 0.51 when tested for P1) which means that the token closely fails the maximum posterior probability criterion and is therefore rejected by agent B. Section 3.3.2.2 is about the long-term effects of the relative criterion.

Memorisation strategies can be combined, i.e. it is possible to apply both the absolute and a relative criterion. Only if both criteria are passed, the token is memorised. Finally, it is also possible to apply no restriction on perception by having the agents memorise all perceived tokens.

3.2.5 Phonological Level

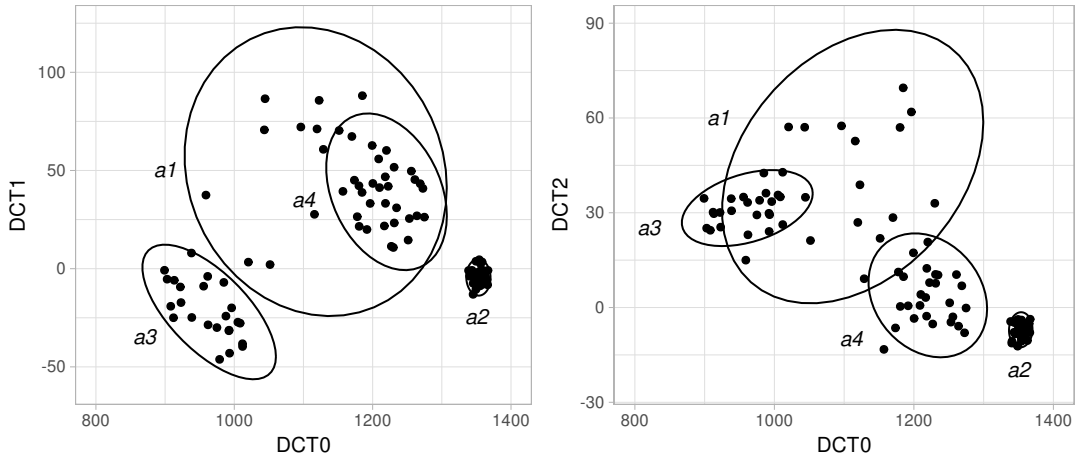
There are two ways of linking the exemplars and word classes through a phonemic level. Either the phonemic classes are fixed and immutable throughout the simulation or they are agent-specific, regularly updated, and computed using unsupervised learning algorithms. The simulations in Harrington and Schiel (2017) are an example of the first option: This is where the user supplies the phonemic labels associated with the word types to the ABM and they are carried along throughout the interactions. Using this fixed phonological level is computationally efficient and may be used if no phonological change is expected to occur. The second option periodically derives phonological knowledge without any prior information and has been shown to be able to model both stability and change on the phonological level (Gubian et al., 2023). That is why it is planned to have the flexible phonology module be the default way of computing the phonological level in the ABM (also see 3.3.1) despite the increased computation time compared to the fixed phonology.

In order to present the two algorithms which are used in the flexible phonology module of the ABM, a simulation was run on the data from Harrington and Schiel (2017) (also see Harrington et al., 2008). The simulation settings are not important here, as only the state of one agent's memory after 100,000 interactions is considered. The agent (called *phfo*) was initialised with 10 exemplars each of 11 word types in a three-dimensional DCT-based acoustic space.

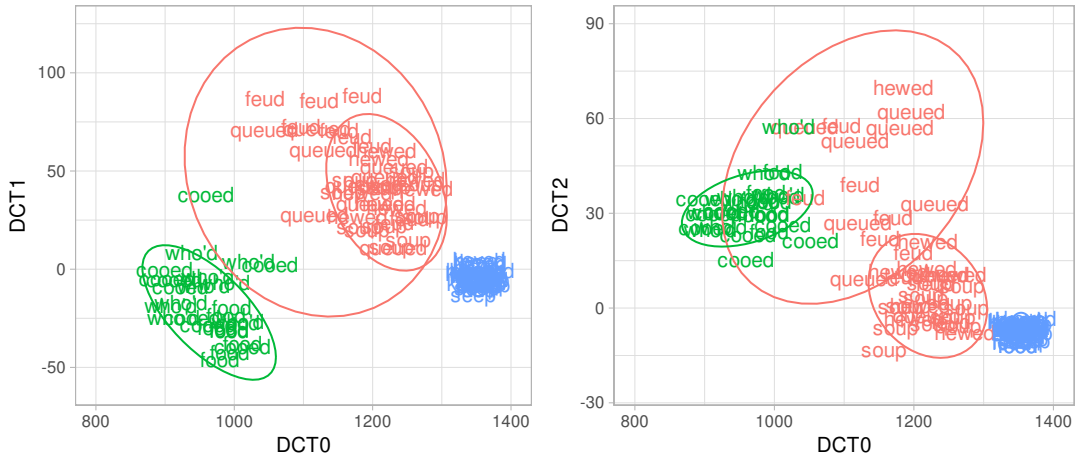
The phonological link between exemplars and words was derived from the following two-step process.

In the first step of the flexible phonology module, Gaussian Mixture Models (GMMs, Reynolds, 2009) are used to create acoustic clusters of exemplars (see Appendix B.4.1 for mathematical details on this algorithm). This step relies exclusively on information about the location of exemplars in the acoustic space and no information about word classes. Figure 3.3a shows the GMM components for the chosen agent at the given point during the simulation. Every black dot represents an exemplar in the agent's memory. At this stage, the exemplars' association to word classes is irrelevant, thus only the acoustic information is used to form clusters in the three-dimensional phonetic space. The GMM has determined in this case that there are four acoustic clusters (*a1* to *a4*) as shown by the labelled ellipses.

The second step is to identify sets of acoustic clusters that contain exemplars of the same distinct word classes. These sets are called sub-phonemes (and not phonemes) for the reasons explained by Gubian et al. (2023). The algorithm which identifies sub-phonemes is non-negative matrix factorisation (NMF, D. D. Lee & Seung, 2001, see Appendix B.4.2 for mathematical details). NMF disregards any information about the location of exemplars and acoustic clusters in the acoustic space, and instead uses information on the association between exemplars and word classes. This is why every exemplar from Figure 3.3a is represented by the corresponding word label in Figure 3.3b. In total, NMF has determined that there are three sub-phonemes as shown by the different colours. Acoustic clusters *a1* and *a4* both contain mostly exemplars of the same four word classes: *queued*, *feud*, *hewed*, and *soup*. That is why *a1* and *a4* are grouped together into the red sub-phoneme. Most exemplars of the words *cooed*, *food*, and *who'd* are contained in acoustic cluster *a3* and no other cluster, hence *a3* becomes the green sub-phoneme. However, two exemplars of *cooed* and one exemplar of *who'd* were originally part of acoustic cluster *a1*. These exemplars are so-called impurities in the red sub-phoneme (with which cluster *a1* is associated). Sub-phonemes can contain impurities as long as the overall purity of the sub-phoneme surpasses a threshold determined by the user. Purity is computed as the fraction of exemplars in a sub-phoneme



(a)



(b)

Figure 3.3: Example of the flexible phonology module given an agent's memory at a certain time during a simulation. The agent has stored exemplars of 11 word types in memory which are first grouped into acoustic clusters *a1* to *a4* by GMM **(a)**. The acoustic clusters are then grouped into three sub-phonemes as indicated by the colour-coding by NMF **(b)**. The three-dimensional DCT-based acoustic space was split into two two-dimensional plots for better clarity.

belonging to a designated set of words. So if the red sub-phoneme consists of a total of 27 exemplars, 24 of which are associated with the word classes *queued*, *feud*, *hewed*, and *soup*, then the purity is $\frac{24}{27} \approx 0.89$. Since this is higher than the default purity threshold of 0.75, *a1* and *a4* are pure enough to become

a sub-phoneme. Finally, acoustic cluster *a2* exclusively contains exemplars of the words *feed*, *heed*, *keyed*, and *seep* and is hence identified as the blue sub-phoneme. A more in-depth explanation of this process is provided in Appendix B.4.3.

Notice that the identified sub-phonemes overwhelmingly correspond to the classical phonemes /i, u, ju/ as indicated by Table 3.1, the only exception being *soup* which was grouped with /ju/- instead of /u/-words by the unsupervised algorithms. Importantly, whenever the sub-phonemic classes are recomputed, any previous results from GMM and NMF are disregarded. If this flexible phonology module is used to derive phonemic knowledge, the memorisation criteria are computed with respect to the agent-specific sub-phonemes.

Phoneme	Word Classes
/i/	feed, heed, keyed, seep
/u/	food, who'd, cooed, soup
/ju/	feud, hewed, queued

Table 3.1: Association between words and classical phonemes (also see Harrington and Schiel, 2017).

3.2.6 Memory Management

There are two scenarios in terms of memory management that should be avoided because they can cause artefacts: data scarcity on the one hand renders the computation of Gaussian distributions unstable, while an abundance of exemplars minimises the impact of new tokens on the exemplar distributions and hence effectively prohibits change from happening. In order to always retain a sufficient number of exemplars per word class in an agent's memory, two measures were implemented. The first counteracts the data scarcity problem during the initialisation of the agents. If there are too few tokens from real speakers available, the Synthetic Minority Over-sampling Technique (SMOTE, Chawla et al., 2002, see Appendix B.5 for mathematical details) can be applied during the initialisation of the agents in order to increase the number of tokens per word and agent. SMOTE is a standard resampling technique that linearly

connects the existing data points to their nearest neighbours and randomly samples new data points on these connecting lines.

The second measure takes effect when the agents' memory size needs to be controlled during the simulation, i.e. when the agents start to accumulate more and more exemplars. Agent listeners are equipped with the ability to forget an exemplar after having accepted and memorised a new exemplar. The exemplar that is removed from memory has to be associated with the same word class as the newly memorised exemplar, but is otherwise chosen at random. However, the removal of the chosen exemplar is blocked if it would lead to a decrease in the number of exemplars of the associated word class beyond the initialisation level. This constraint was implemented to prevent the agents from forgetting entire word classes. It is the user's task to adjust the rate of exemplar removal on a spectrum between no removal at all to removing an exemplar every time a new exemplar has been memorised.

3.2.7 From Interactions to Change

According to the IP model (see section 3.1), permanent sound change can arise when an existing phonetic bias is reinforced by interactions between individuals who imitate each other and whose associations between word classes and remembered speech signals can be updated with increasing language experience. That is, there are four key components in the IP model whose interplay determines whether or not a sound change takes place: phonetic biases, non-social phonetic imitation, perceptual learning, and flexible subphonological categories. This section is concerned with the mapping between these theoretical components and their counterparts in the ABM in order to understand under which circumstances the simulations can result in change.

The first component to be discussed here is phonetic bias which results from how speech is produced or perceived. In the IP model, the directional synchronic variation that is a consequence of stable phonetic biases is a precondition for sound change to emerge. In terms of the ABM, this highlights the importance of the input data, i.e. the data configuration at baseline determines in no small part the simulation's outcome. In particular, the input

data should mirror some phonetic variation that may eventually turn into sound change. This is important because the mechanisms of the ABM are incapable of generating systematic variation or a phonetic bias. As explained in section 3.2.3, the agents' production is a sampling procedure that is not subject to a superimposed bias which would skew the produced token in the direction of the expected sound change. The way in which variation is injected into the model is by initialising agents with data from heterogeneous (groups of) speakers. Some of these speakers should be further along in the modelled sound change than the others such that, when speakers with different variants interact, it can be observed whether all of them end up producing the more innovative variant.

In the IP model, phonetic imitation, i.e. the acoustic convergence between (usually two) speakers in a conversation, is viewed as an involuntary, subconscious pattern of behaviour which is not socially motivated. The same is true for the ABM: agents propagate their variants of a sound not because they are instructed to build a common identity nor because they desire to become a member of a social group, but because of the strong link between their speech perception and speech production. That is, both processes draw from the same pool of exemplars and newly memorised exemplars can affect the agent's production. The agents' production-perception feedback loop is what can turn phonetic variation into population-wide change in the computational model.

The third component of sound change in the IP model is perceptual learning, i.e. the possibility to shift perceptual boundaries between phonological categories upon encountering ambiguous tokens of words. When the phonological classes in the ABM are pre-determined and remain fixed, perceptual learning is more easily observed than when they are derived automatically from the acoustic exemplars. For example, in the scenario shown in Figure 3.2, two agents have different representations of phoneme P1 but essentially the same representation of phoneme P2. When either of them produces a token of a word that is associated with P1, it is likely going to sound unfamiliar or ambiguous to the perceiving agent. After a prolonged exchange of exemplars of P1, agent A is more prone than agent B to shifting the perceptual boundary between the phonemes P1 and P2 towards P2 because agent A's P1 is skewed

towards P2. However, the ambiguous exemplars still have to pass the relative memorisation criterion, otherwise perceptual learning cannot take place. This is where the flexible phonology module (see section 3.2.5) might be helpful since the parts of the acoustic space considered ambiguous by individual agents can change each time the sub-phonemic classes are recomputed. In this case, perceptual learning might be understood as the creation of phonological or sub-phonemic classes that reflect the agent's acquired knowledge about the association between speech signals and word classes which is updated when the agent has collected more language experience (Cutler et al., 2010). So perceptual learning is also closely related to the idea of a flexible and personalised phonological level, i.e. the fourth component of the IP model.

Recall that the IP model states that "whether or not sound change actually comes about depends upon which speakers regularly speak to each other, whether a phonetic bias happens to be magnified by interaction, and whether or not sub-phonemic classes are fragmented and regrouped over time" (Harrington et al., 2018, p. 17). In the ABM, sound change can emerge from the stochastic interactions between heterogeneous agents (some of which are initialised with a more innovative variant of the sound under investigation than others), given the mechanics of their production-perception feedback loop and organisation of phonological information. The components of the theoretical IP model and computational ABM described here also fit the definition of spoken language as a complex adaptive system according to Beckner et al. (2009) (section 3.1): multiple members of a speech community (i.e. speakers or agents) interact with one another; their interactions are governed by the mechanics of speech production and perception which can be adapted as a consequence of past interactions; and sound change arises from the interplay between language experience, population dynamics, and cognitive processes.

3.3 Core Mechanisms

The ABM presented in this chapter offers a wide variety of setting combinations which, also in combination with the endless possibilities of input data, can yield wildly differing results. In this section, we focus on two of the essen-

tial mechanisms of this model and how they affect certain artificially generated data configurations. Firstly, the simulations in 3.3.1 show that it is advantageous to use the flexible as compared to the fixed phonology module in order to establish phonological classes as a link between exemplars and word classes. Secondly, three simulations are presented in 3.3.2 that explore the effects of the absolute and relative memorisation criterion on the artificial dataset first shown in Figure 3.2 in order to show how simulations can result in both phonetic change and stability.

3.3.1 Flexible vs. Fixed Phonology

The aim of this section is to demonstrate the main advantage of using the flexible phonology module that was presented in section 3.2.5: The phonological level is derived directly from the data without relying on any superimposed information provided by the user. Two artificial datasets were created to show the implications of this data-based approach to generating phonological knowledge. Both consist of only one agent who is initialised with 10 exemplars each of 10 word classes, called W1 to W10, in a two-dimensional acoustic space. The only difference between the two datasets is the association between exemplars and word classes which is either systematic or random. Without conducting any interactions, the flexible phonology module is used to identify sub-phonemic classes in this data.

Figure 3.4 shows the first dataset. All exemplars are represented by their word class in the two-dimensional acoustic space. The Gaussian Mixture Model has identified two acoustic clusters, *a1* and *a2*. Since *a1* exclusively contains exemplars of words W1 to W5 and *a2* overwhelmingly contains exemplars of words W6 to W10, each acoustic cluster constitutes a sub-phoneme on their own according to NMF (as indicated by the ellipses of different line types). One exemplar of W2 is part of *a2* instead of *a1* (which holds all other exemplars of W2), so this exemplar is considered an impurity in sub-phoneme 2 (SP2, dashed ellipse). However, SP2's overall purity is still very high with a value of $\frac{50}{51} \approx 0.98$. From the example in Figure 3.4 it can be concluded that the machine learning algorithms that were applied to compute the phonological

level were capable of recognising the systematicity in how exemplars were associated with word classes: namely that exemplars of word classes W1 to W5 are usually characterised by negative values of Feature 1 whereas exemplars of W6 to W10 have positive values of Feature 1.

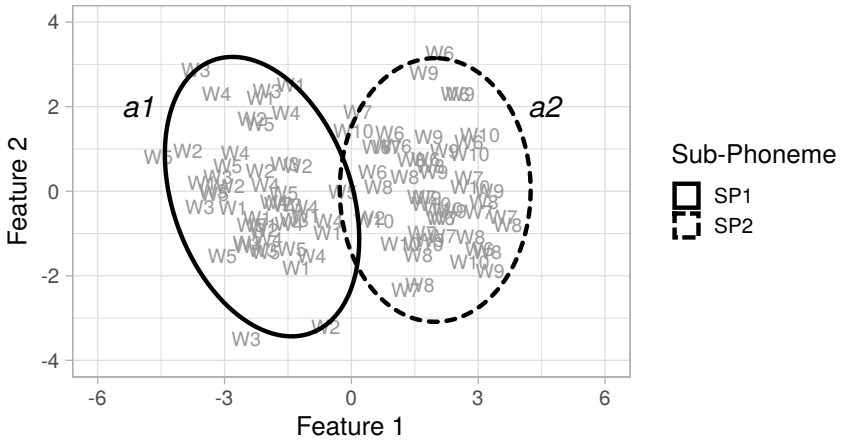


Figure 3.4: Results of applying GMM and NMF to an agent whose association between exemplars and word classes is systematic. Each exemplar is represented by its corresponding word class in the two-dimensional acoustic space. The flexible phonology module identified two sub-phonemes (SP1 and SP2, as indicated by the ellipses), each consisting of one acoustic cluster, *a1* and *a2*, respectively.

The position of exemplars in the acoustic space is the same in the second as in the first dataset. In this case, however, the word labels were randomly assigned to the exemplars. Figure 3.5 shows that a mere swapping of word labels in comparison to the first dataset yields fairly different results. GMM and NMF have determined that there is only one sub-phoneme which consists of two acoustic clusters. These two acoustic clusters are the same as in Figure 3.4 because the location of exemplars in the acoustic space is the same. However, since both *a1* and *a2* contain exemplars of the same (i.e. all) word classes, NMF has grouped the two clusters together to form one sub-phonemic class. That is, the flexible phonology module has correctly inferred that there was no discernible phonological pattern in this data.

The scenarios in this section have shown that the flexible phonology module is capable of drawing reasonable conclusions from both systematically and

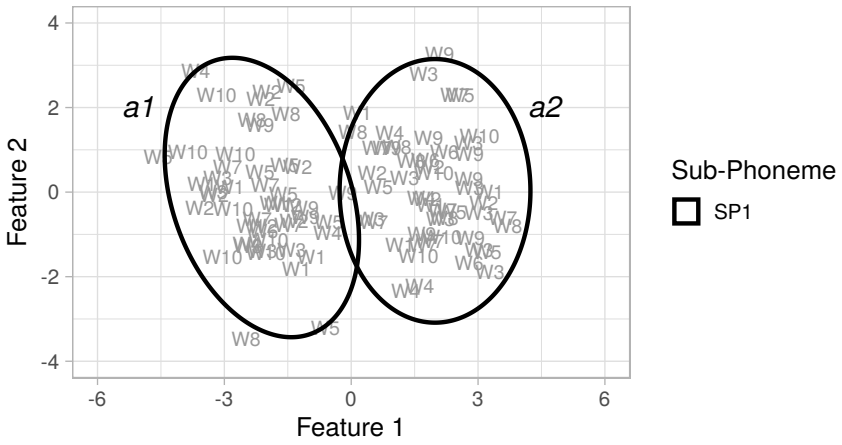


Figure 3.5: Results of applying GMM and NMF to an agent whose association between exemplars and word classes is random. Each exemplar is represented by its corresponding word class in the two-dimensional acoustic space. The flexible phonology module identified one sub-phoneme (SP1) consisting of two acoustic clusters (*a1* and *a2*).

randomly distributed data. In these two cases, where no interactions took place, a user possibly would have come to the same conclusions, i.e. that in one case there should be two phonological classes and in the other only one. However, the main advantage of the flexible as compared to the fixed phonology module emerges more clearly when the agent's exemplar storage grows and changes as a result of interactions with other agents. While GMM and NMF can react to an agent's increasing language experience and adapt the phonological link between exemplars and words accordingly, it is impossible for the user to intervene and determine new phonological classes during a simulation. When the flexible phonology module is used, sub-phonemes are directly derived from data, i.e. in accordance with Exemplar Theory they are abstractions over clouds of stored exemplars (Pierrehumbert, 2001), thereby eliminating the risk of superimposing fixed phonological classes that do not (anymore) match the acoustic data. This is important because, as mentioned in section 3.2.4, the memorisation criteria which are one of the main forces in the ABM are computed with regard to the (sub-)phonological classes. Finally, GMM and NMF allow for phonological changes such as splits and mergers to happen. It is

beyond the scope of this chapter to demonstrate phonological changes, but the reader is referred to Gubian et al. (2023), for a demonstration of phonological stability and change in this agent-based model.

3.3.2 Phonetic Stability and Change

This section is concerned with showing the long-term effects of the absolute and relative memorisation criteria on the outcome of simulations using the artificial data that was shown in Figure 3.2. The data consists of two agents, A and B, both of which are initialised with 50 exemplars each of 10 word classes. The acoustic space is two-dimensional and the phonological level was determined using GMM and NMF as explained in section 3.2.5. The data was intentionally designed in such a way that the initial computation of the phonological level would yield two sub-phonemic classes for each agent.[7] For agent A, sub-phoneme 1 (SP1) is broad and skewed towards sub-phoneme 2 (SP2) which, in turn, is similar to agent B's SP2. For agent B, both sub-phonemic classes are relatively narrow and in close proximity in the acoustic space. This setup is motivated by the same findings that support the IP model (section 3.1). That is, synchronic phonetic variability can skew a phonological class (in this case agent A's SP1), e.g. in many English varieties /u/ is fronted in adjacency to tongue-tip or palatal consonants, but remains a high back vowel in other contexts. In another group of speakers (represented here by agent B), however, the same phonological category is less variable and might already lie in the direction of change, e.g. when some speakers have already adopted a fronted /u/ in all contexts.

Below, three simulations are presented which use the described artificial data as input. The first demonstrates the impact of the absolute memorisation criterion, while the second shows the effect of applying the maximum *a posteriori* decision in perception (see section 3.2.4). The third applies both criteria, as is the default in the `soundChangeR` settings. The simulations were stopped

[7] Sub-phonemic labels are randomly assigned at every rerun of GMM and NMF and they are also not the same across agents. For reasons of convenience, the labels of the sub-phonemic classes in Figures 3.6 to 3.10 have been manipulated to be the same before and after the simulation.

when there were no more acoustic changes (after 5800, 9000, and 6000 interactions respectively in the first, second, and third simulation). Apart from that, the simulations share the same settings: the phonological level was computed using GMM and NMF and recomputed whenever 20 new exemplars had been memorised; sub-phonemic classes had to surpass a purity level of 75%; and after each memorisation, another exemplar was removed from memory.

3.3.2.1 Effects of the Absolute Memorisation Criterion

Figure 3.6 shows the acoustic space of the two agents (in columns) at the baseline (top row) and post-run (bottom row). Every grey dot represents a memorised exemplar and the ellipses indicate their association to sub-phonemic classes SP1 (solid) and SP2 (dashed). The Mahalanobis distance threshold was set to 0.95 probability mass (the default setting), which means that an exemplar is accepted and memorised only if it is located within the mean-centred ellipse enclosing 95% of the probability mass of the corresponding Gaussian distribution. From the bottom row of Figure 3.6 it can be seen that after the interactions both agent A and agent B still have two sub-phonemic classes, each of which consists of one acoustic cluster. It is also very clear that agent A has adopted agent B's variant of SP1 so that the two sub-phonemes are in close proximity and even overlap in the acoustic space. Over simulation time, the agents also forget a number of exemplars, so agent A has removed many of the original exemplars of SP1 with low values of Features 1 and 2, leaving A with a shifted and narrower SP1. SP2, on the other hand, has not changed in either agent. Also note that the agents' sub-phonemic classes have generally become more narrow post-run compared to the baseline.

Insights on how these simulation results came about can be taken from Figure 3.7 which shows the baseline configuration for both agents (grey dots and ellipses) as well as exemplars that were rejected (yellow) or accepted (blue) by the absolute criterion in the first 200 interactions when tested for membership in SP1 (top) or SP2 (bottom). Exemplars of SP1 produced by agent A typically have low values of Feature 1 and mostly low values of Feature 2. These exemplars would be very atypical members of agent B's SP1; in technical

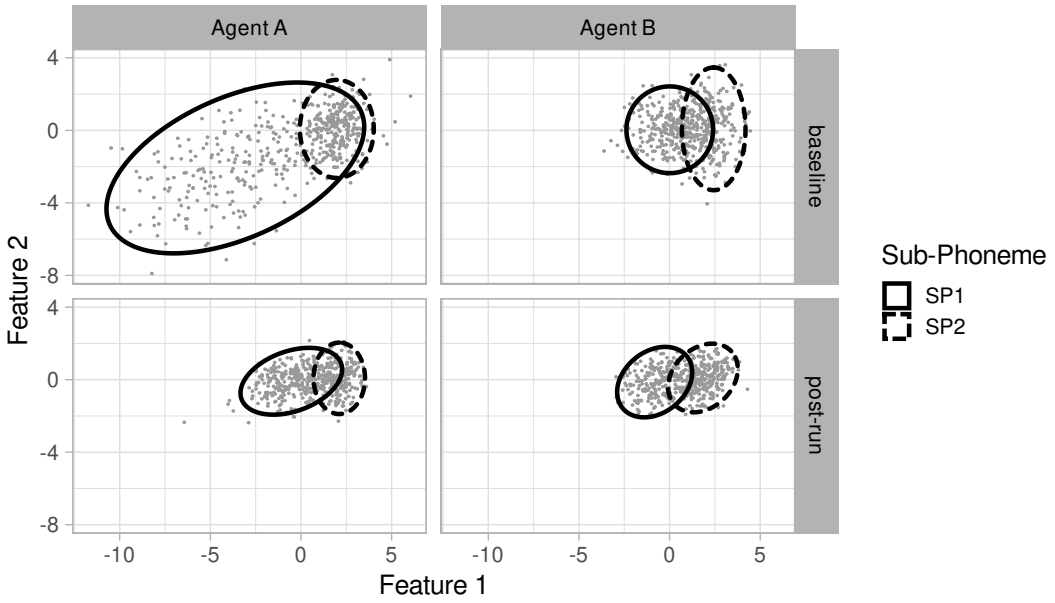


Figure 3.6: State of agent A's and B's representations of sub-phonemic classes in the acoustic space before and after the simulation. Grey dots represent stored exemplars. Only the absolute Mahalanobis distance criterion with the default threshold of 0.95 was applied to decide whether perceived tokens should be memorised.

terms, that means that these exemplars' Mahalanobis distance to the centroid of agent B's SP1 is too high and they are therefore rejected (cf. yellow dots in top right panel). As a consequence, agent B's SP1 does not shift towards that of agent A. The exemplars of SP1 produced by agent B, however, all fall within agent A's ellipse and thus are accepted according to the absolute criterion and memorised (cf. blue dots in top left panel). Since all of the exemplars that agent A accepts for SP1 lie in the top right corner of the acoustic space, agent A's SP1 starts to shift in that direction. So because agent A's SP1 is large and skewed towards agent B's narrow SP1, agent A is more likely to accept tokens from agent B than vice versa, resulting in a shift of agent A's SP1 towards that of agent B. SP2, on the other hand, is very similar in both agents, so most of the produced tokens pass the Mahalanobis distance threshold (blue dots in bottom row). The rare tokens that are rejected (yellow dots in bottom row) closely miss the threshold. This rigid pressure against atypical tokens also leads to ever

narrower sub-phonemic classes, because even exemplars that are almost, but not quite enclosed by the ellipses are rejected.

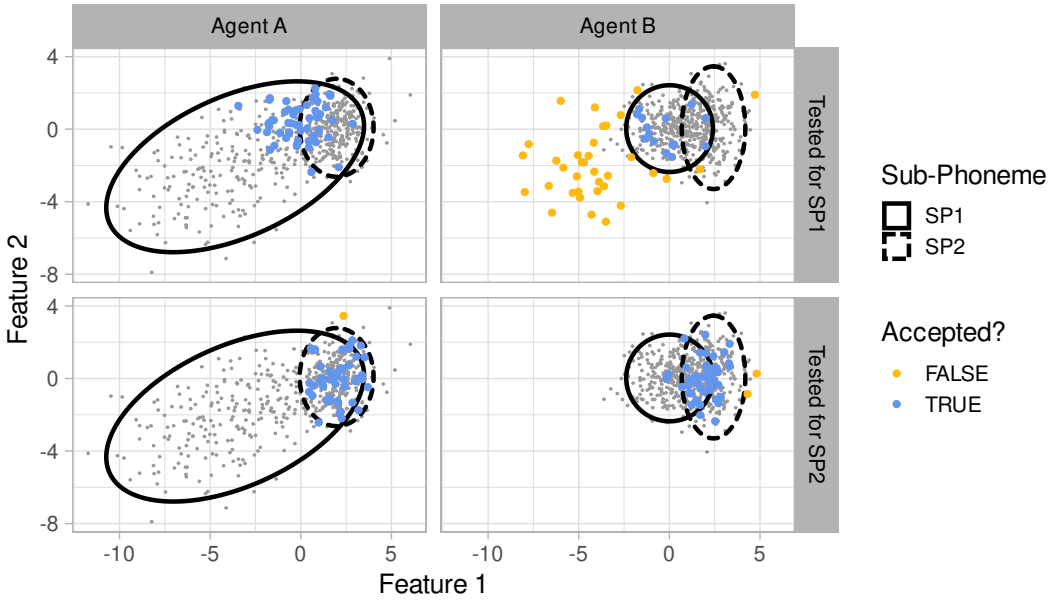


Figure 3.7: Baseline configuration of sub-phonemic classes 1 (SP1, solid) and 2 (SP2, dashed) for agents A and B (in columns) in the acoustic space. Grey dots represent baseline exemplars, yellow/blue dots represent exemplars from the first 200 interactions that failed/passed the Mahalanobis threshold when tested for membership in SP1 (top) or SP2 (bottom).

3.3.2.2 Effects of the Relative Memorisation Criterion

Similarly to Figure 3.6, Figure 3.8 shows the sub-phonemic classes 1 (solid ellipse) and 2 (dashed ellipse) of the two agents (in columns) both before (top row) and after the simulation (bottom row). In this case, the relative criterion (i.e. maximum *a posteriori* probability) was used to decide whether or not a token was stored in an agent's memory. After 9000 interactions between agents A and B, agent B's SP1 has shifted towards that of agent A and both agents' phonological classes are acoustically distinct.

As shown by Figure 3.9, these results are a consequence of a pressure against ambiguous exemplars, i.e. exemplars that are located in a part of the

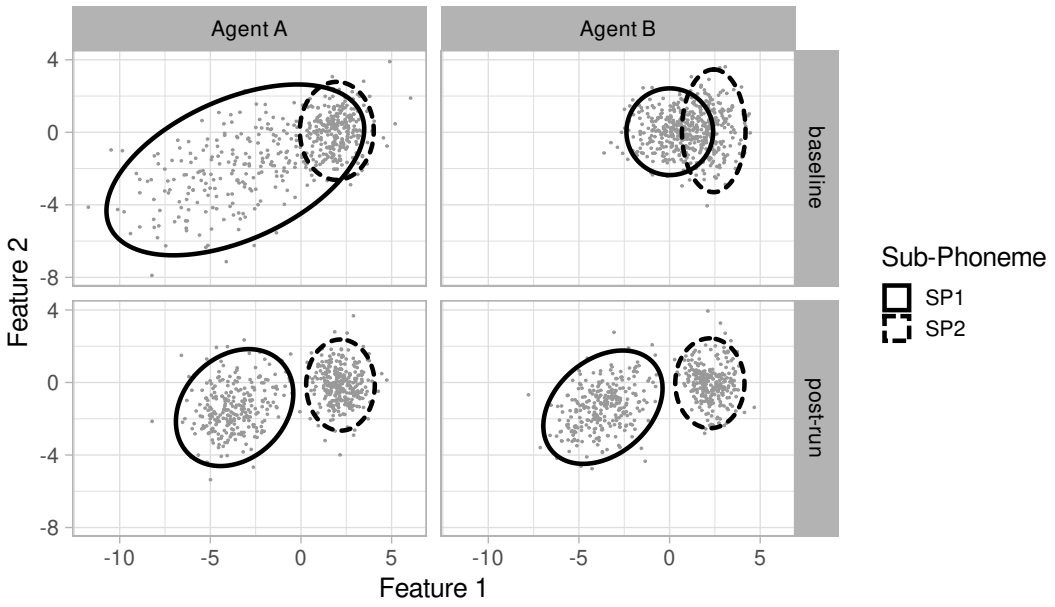


Figure 3.8: State of agent A's and B's representations of sub-phonemic classes in the acoustic space before and after the simulation. Grey dots represent stored exemplars. Only the relative maximum posterior probability criterion was applied to decide whether perceived tokens should be memorised.

acoustic space in which it is questionable according to posterior probabilities whether the exemplar belongs to one or another sub-phoneme. For instance, agent B is very likely to accept exemplars of SP1 from agent A since these almost always fall into an unambiguous part of the acoustic space, i.e. a part where SP2 cannot be considered a competing phonological class according to the conditional posterior probabilities (cf. top right panel). Agent A, on the other hand, rejects those exemplars of SP1 produced by agent B that are located in the part of the acoustic space where agent A's SP1 and SP2 overlap (yellow dots in top left panel). From the top row of Figure 3.9 it also becomes clear that agent B's SP1 must have shifted towards that of agent A already after 200 interactions, given that agent B has started to produce exemplars of SP1 that are unambiguous to agent A and therefore accepted (blue dots in top left panel). Similarly to the first simulation, both agents usually accept each other's exemplars of SP2, because they have acoustically similar representations of that sub-phonemic class. Interestingly, however, agent B has rejected several

exemplars of SP2 because they fell into the part of the space where SP1 and SP2 overlap (yellow dots in bottom right panel). Overlapping sub-phonemes cause ambiguity, and such ambiguity is penalised when the relative criterion is applied. Over time, the constant pressure against ambiguous exemplars creates a force of repulsion between sub-phonemes as can be seen from the bottom row of Figure 3.8 where the agents acoustically discriminate SP1 and SP2.[8] Notably, some of the exemplars accepted by both agents are clear outliers of their sub-phonemic classes (at least as long as the sub-phonemic classes have not been adapted by reapplying GMM and NMF), i.e. they are not enclosed by the ellipses (also see Figure 3.8). A side effect of applying the relative memorisation criterion is that outliers are accepted as long as they are located in an unambiguous part of the acoustic space, which can lead to an expansion of the acoustic space even beyond values that are realistic in terms of articulation.

In summary, the two simulations in this section have both shown cases of phonetic change and stability. When agents did not differ in their representation of a sub-phonemic class (e.g. SP2), there was no phonetic change, whereas between-agent variation (e.g. in SP1) in combination with the different perceptual strategies resulted in shifts of sub-phonemic classes in the acoustic space as well as in changes of the classes' size and orientation. This underlines the necessity for heterogeneous agents or agent groups: the acoustic variation introduced by them is required for phonetic changes to emerge from a simulation with the agent-based model. It was also shown that the two memorisation criteria had opposing effects on the chosen data configuration: the Mahalanobis distance threshold caused a shift from agent A's SP1 towards that of agent B, whereas the maximum *a posteriori* probability resulted in a shift from agent B's SP1 towards that of agent A. Since both memorisation criteria have different long-term effects as well as possible side effects, it seems reasonable overall to test exemplars both for their typicality by means of a Mahalanobis distance

[8] Recall, however, that sub-phonemes are recalculated regularly, which can lever out the effect of repulsion caused by the application of the relative criterion. For further explanation, the reader is referred to Stevens et al. (2019) and Gubian et al. (2023).

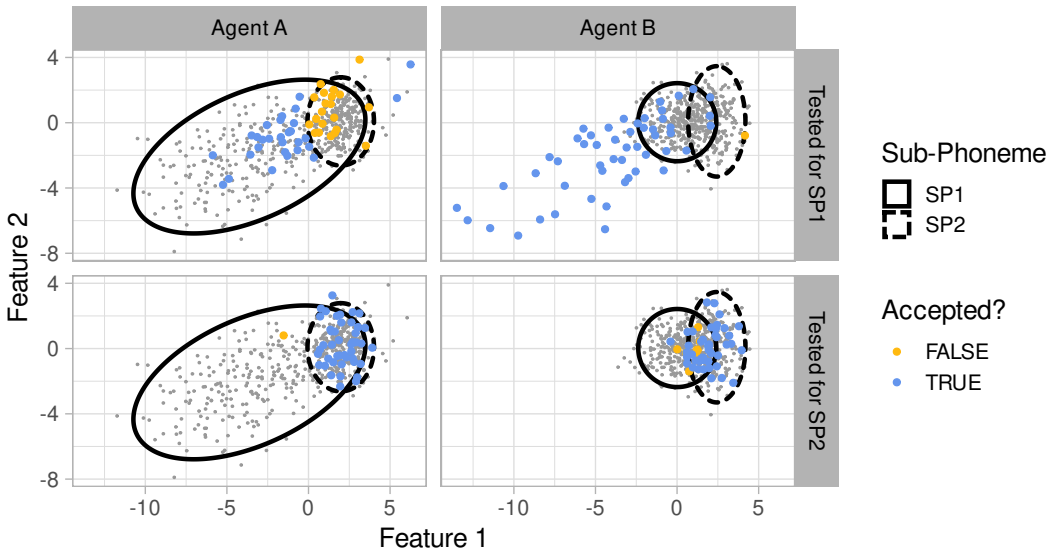


Figure 3.9: Baseline configuration of sub-phonemic classes 1 (SP1, solid) and 2 (SP2, dashed) for agents A and B (in columns) in the acoustic space. Grey dots represent baseline exemplars, yellow/blue dots represent exemplars from the first 200 interactions that failed/passed the maximum posterior probability decision when tested for membership in SP1 (top) or SP2 (bottom).

threshold and their discriminability by means of posterior probabilities (Todd et al., 2019).

3.3.2.3 Effects of Applying Both Memorisation Criteria

Figure 3.10 shows a simulation with the same data as before, but when both criteria had to be passed in order for a token to be memorised. Both agents have maintained two sub-phonemic classes consisting of one acoustic cluster each. Agent A's SP1 has shifted more towards agent B's SP1 than vice versa, the sub-phonemic classes do not overlap anymore, and they have become slightly narrower over simulation time. So the simulation results display characteristics of both of the previous simulations.

Figure 3.11 illustrates how this outcome arose as a consequence of the application of both memorisation criteria. All exemplars that were exchanged in the first 400 interactions between the agents are plotted over the baseline

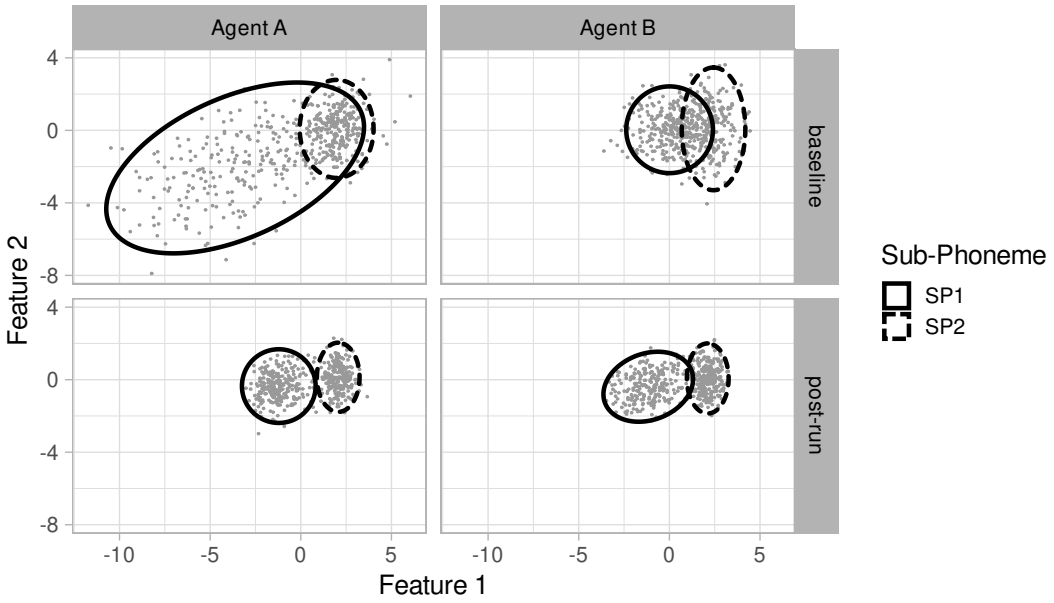


Figure 3.10: State of agent A's and B's representations of sub-phonemic classes in the acoustic space before and after the simulation. Grey dots represent stored exemplars. Both memorisation criteria had to be passed during the perception procedure in order for a perceived token to be memorised.

configuration of the data with the colour-code indicating whether the exemplar was accepted (blue), rejected due to the absolute (yellow) or relative criterion (red). More precisely, exemplars were tested first against the relative criterion; only if they passed that test, they were also subjected to the absolute criterion. So yellow exemplars in Figure 3.11 were accepted according to the relative criterion, but rejected by the absolute criterion, whereas red exemplars were rejected by the relative criterion even though they might have passed the absolute criterion if they had been subjected to it.

Agent A rejected exemplars of SP1 produced by agent B because they were too probabilistically close to SP2 (red dots in top left panel). Just like in section 3.3.2.1, agent A was more inclined to accept exemplars of SP1 produced by agent B (blue dots) than vice versa because agent A's SP1 is broad and skewed in the direction of agent B's SP1. Exemplars of SP1 produced by agent A, while clearly acceptable to agent B according to the relative criterion (cf. Figure 3.8), were rejected because of the Mahalanobis distance threshold (see yellow dots

in top right panel of Figure 3.11, cf. Figure 3.7). That also means that the effects of the maximum posterior probability decision – i.e. an expansion of the acoustic space through the acceptance of outliers in unambiguous parts of the acoustic space – was counteracted by the absolute criterion and resulted in only minor phonetic changes of agent B's SP1.

With regard to SP2, Figure 3.11 shows that the agents almost always accepted and memorised each other's tokens (blue dots, bottom row) because they share a similar representation of that sub-phonemic class. Agent A rejected some of agent B's exemplars of SP2 because they did not pass the Mahalanobis distance threshold despite their proximity to agent A's SP2 (yellow dots in bottom left panel) and others because they were probabilistically closer to agent A's SP1 and SP2 (red dots). Finally, as shown also in Figure 3.9, agent B rejected some exemplars of SP2 produced by agent A because they were too ambiguous to classify according to the maximum posterior probability decision (red dots in bottom right panel of Figure 3.11). Since the agents' sub-phonemic classes are relatively stable over simulation time, the relative criterion can create the force of repulsion mentioned in section 3.3.2.2 which is why SP1 keeps its distance to SP2 (cf. bottom row of Figure 3.10).

This simulation, in which both memorisation criteria had to be passed, demonstrates the interplay between the two and their combined effect on the chosen data configuration: The absolute criterion is capable of enforcing phonetic skews if the agent with the skewed variant is in contact with another agent whose variant lies in the direction of the skew. Thereby (and in combination with a rigorous removal policy), the absolute criterion can create shifts of (sub-)phonemic classes through the acoustic space. By contrast, the relative criterion takes effect when two (sub-)phonemic classes are acoustically close and prohibits the memorisation of ambiguous exemplars. As a consequence, (sub-)phonemic classes tend to stay apart.

3.4 Discussion

This chapter was about a cognitively-inspired agent-based model of sound change: its theoretical background based on the IP model, its entities and pro-

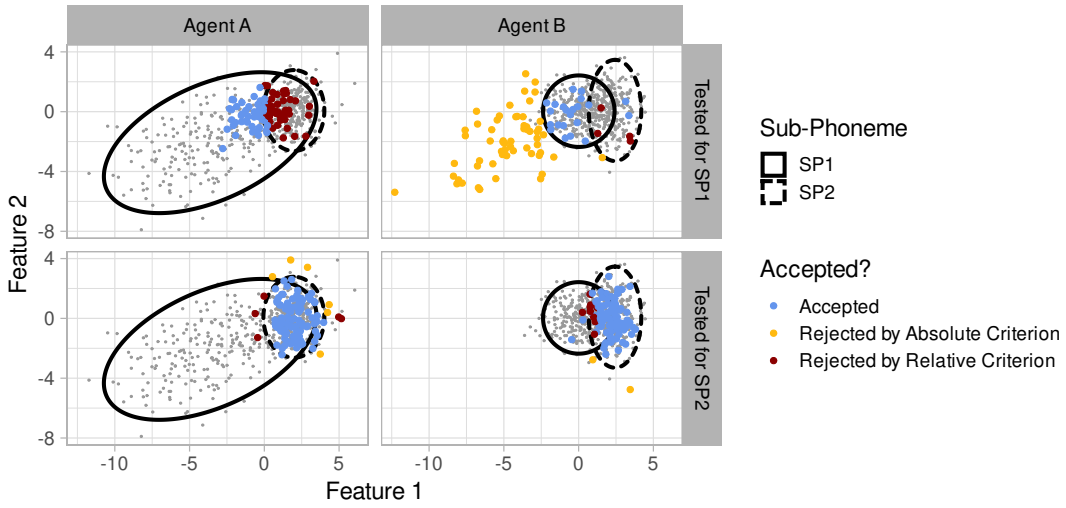


Figure 3.11: Baseline configuration of sub-phonemic classes 1 (SP1, solid) and 2 (SP2, dashed) for agents A and B (in columns) in the acoustic space. Grey dots represent baseline exemplars, blue dots represent exemplars from the first 400 interactions that passed both criteria, yellow/red dots represent exemplars that were rejected by the absolute/relative memorisation criterion when tested for membership in SP1 (top) or SP2 (bottom).

cesses, the advantages of using the flexible as compared to the fixed phonology module, and how all of this together can result in change. Here we aim to compare the properties of the ABM to those of other computational models of sound change, point out some unresolved issues, and discuss opportunities to expand the model.

Compatibly with many other agent-based or computational models of sound change (Blevins & Wedel, 2009; Ettlinger, 2007; Kirby, 2014b; Kirby & Sonderegger, 2013; Pierrehumbert et al., 2014; Sóskuthy, 2015; Stanford & Kenny, 2013; Todd et al., 2019; Wedel, 2006), `soundChangeR` is founded upon an exemplar-based production-perception feedback loop as proposed by exemplar theory and episodic models of memory. That is, the agents store parameterised traces of speech in their memories, are capable of deducing some form of phonological code from the phonetic and lexical levels, and use their knowledge to produce new exemplars and to perceptually evaluate and categorise incoming tokens. In contrast to all other computational models of

sound change, the ABM presented in this chapter can handle real production data and is not reliant on artificially created data. While artificial data can be helpful in exploring the simulations' possible outcomes by controlling their input (as demonstrated in section 3.3), artificial data can never adequately include all the idiosyncrasies of spoken language which have been shown to be relevant for sound change (Beddor et al., 2018; Clopper, 2014; Harrington, 2014; Stevens & Harrington, 2014; Yu, 2021; Yu & Zellou, 2019). The exemplars in other ABMs follow either a uniform distribution (Wedel, 2006) or normal distribution in a finite artificial acoustic space, e.g. between 0 and 1 (Sóskuthy, 2015) or between 0 and 30 (Stanford & Kenny, 2013). In the models by Kirby (2013, 2014b) the acoustic features are distributional information derived from real production data, i.e. the agents do not store single exemplars but rather retain information about the acoustic distribution. The production data with which agents are initialised in `soundChangeR` should mirror phonetic variation as well as different stages of the sound change under investigation, i.e. some agents should be further advanced in the change than others. This is often achieved by means of apparent-time data in which it is assumed that younger speakers represent a more recent state of the language than older speakers (Bailey et al., 1991). In interactions between such heterogeneous agents, it is expected that all agents adopt the innovative variant of the sound by the end of the simulation. However, such a result should not be taken to mean that older speakers adapt to younger speakers in everyday life and that this is how sound change comes about – it only shows that a newer state of the language acts as an attractor to people who use an older speaking style while older states of the language are abandoned.

Although most of the computational models cited above rely on the same theoretical background, they differ in their precise implementation and application of production and perception which, as shown in section 3.3.2, can greatly influence the simulation's outcome. Focusing first on perception, we can compare how different ABMs impose rules and constraints on the categorisation of exemplars. For example, the model by Stanford and Kenny (2013) only applies a relative criterion, so the perceived exemplar has to be closer to the intended than to competing phonological classes in order to be memorised.

Similarly, the models by Kirby (2013, 2014b) use an ideal observer in perception, i.e. a Bayesian classifier that relies on posterior probabilities in order to decide whether an incoming token is accepted. Sóskuthy (2015) weights the relative memorisation criterion by functional load, i.e. ambiguous exemplars are likely to be rejected unless their functional load is low. Common to all of these models is their focus on contrast maintenance. Thus, when there are two or more competing phonological categories in the simulated acoustic space, tokens that fall in the ambiguous space between them are less likely to be memorised than those that are probabilistically close to their intended phoneme. This perceptual pressure for contrast maintenance effectively prohibits mergers from happening in these models because phonemes will always repel each other. Gordon (2013) notes that chain shifts and mergers are alternative outcomes of phonetic changes whereby one phoneme approaches another in an acoustic space: If the phonemic contrast must be maintained (e.g. in order to avoid wide-spread cases of homophony in the lexicon; Blevins and Wedel, 2009), a chain shift occurs; otherwise the two categories merge and the contrast is lost. Other computational models of sound change apply no restrictions or constraints on the agents' perceptual procedure (Fagyal et al., 2010; Lev-Ari, 2018), but this is because they are interested in the influence of social network size on the propagation of phonetic variants and less so in modelling the cognitive mechanisms inherent in human speech perception. The only ABM besides `soundChangeR` to my knowledge that applies both an absolute and a relative criterion in perception is that by Todd et al. (2019). According to them, a computational model of phonetic change must meet two desiderata when agents are initialised with two categories. First, the categories should maintain their distance to one another, and second, they should maintain their degree of overlap. The forces needed to fulfil these requirements can come from the application of the two different perceptual criteria, as shown in section 3.3.2: the absolute criterion can lead to a phonetic shift while preserving the categories' overlap whereas the relative criterion promotes contrast maintenance by penalising acoustic ambiguity. Todd et al. (2019) justify these desiderata using data from the New Zealand English vowel shift from Hay et al. (2015) which shows that the front vowel categories /æ/ and /ɛ/ shifted phonetically over the

course of 150 years but maintained both their distance and overlap. Applying both memorisation criteria is indeed apt for modelling phonetic shifts, but fails at modelling changes with phonological components such as mergers or splits because of the rigid definition of phonological categories.

While the assumption that acoustically close phoneme categories do not always collapse is generally reasonable (Blevins & Wedel, 2009), a holistic model of sound change should be capable of producing both stable phonemic conditions as well as mergers and splits. In order to allow categories to merge, agents must be able to update their association between acoustic tokens and phonemic categories regularly so as to weaken the effects of the contrast maintenance rule. More specifically, when an agent has acquired a number of new exemplars through interactions with others, the agent should determine whether its phonological categories still match the acoustic data stored in its memory, and if not, recompute them (Pierrehumbert, 2001). The first computational model of sound change that included a more flexible phonology module is the one by Stevens et al. (2019) (an early version of `soundChangeR`) which was built to simulate the retraction of /s/ in /str/-clusters in Australian English (also see Harrington et al., 2018). In this model, the agents were initialised with real speech data from Australian English speakers producing words containing /s/ (e.g. *seen*, *stream*) or /ʃ/ (e.g. *sheep*). During the simulation, the agents could perform binary splits and mergers of the pre-defined canonical phonemes /s/ and /ʃ/ as a result of which they quickly developed their own, personalised sub-phonemic classes. An analysis of those exemplars which contained /str/ over the whole agent population showed a slightly decreased first spectral moment and thus a more /ʃ/-like realisation of the sibilant in /str/ after as compared to before the simulation. In addition, the phonemic categorisation of these exemplars had changed. After the simulation they were more often separated from /s/ and more often merged with /ʃ/, or they were categorised as their own phonemic class separate from both /s/ and /ʃ/. Importantly, the recomputation of phonological classes throughout the simulation counteracted the effects of the relative criterion that would otherwise lead to phoneme repulsion. However, this approach by Stevens et al. (2019) has two main disadvantages. First, the user has to provide the model with prior knowledge

on the canonical phoneme categories. This means that the categories which are superimposed on the input data do not necessarily fit the actual acoustic distributions, but rather match classical phonemes (Ladd, 2006). In exemplar theory – from which the ABM at hand borrows many principles as explained in section 3.1 – phonological classes are derived bottom-up from clouds of stored exemplars which are different for each person given their individual language experience, and not top-down by relying on minimal pairs in the lexicon to identify distinctive segments (Pierrehumbert, 2001, 2003a). Second, the binary splits and mergers in Stevens et al. (2019) result in very high number of sub-phonemic classes, some of which only comprise of exemplars of two word classes (because this limit was implemented into the model). This is because the algorithm preferred splits over mergers, leading to a fragmentation of the agents' memories. The idea of allowing agents to re-evaluate their internal phonological structures after acquiring new exemplars and tolerating sub-phonemic units which are not necessarily distinctive like classical phonemes was carried over into `soundChangeR`. Two machine learning algorithms are regularly applied to the agents' memories which group the stored exemplars into acoustic clusters (i.e. Gaussian mixture components) and then determine sets of clusters which contain the same distinct subset of word classes. These algorithms are classified as unsupervised, i.e. they do not require any prior information about e.g. the amount of (sub-)phonological classes, and they do not take results from previous computations into account. That is, as shown in section 3.3.1, these algorithms can infer agent-specific sub-phonemic classes solely from the location of exemplars in the acoustic space as well as from their association with word classes. This flexible phonology module was tested by Gubian et al. (2023) on the merger of /ɪə, eə/ in New Zealand English (NZE) in which the first element of /eə/ was raised so much that the formerly distinct lexical sets NEAR and SQUARE are now produced with the same diphthong /ɪə/. The agents in this study were initialised with exemplars of the two diphthongs from older and younger NZE speakers who represented different stages of this sound change; older speakers still produced acoustically distinct centring diphthongs while younger speakers had completed the sound change and hence did not differentiate between /ɪə/ and /eə/ anymore. The simulation was run

on a pre-publication version of `soundChangeR`, i.e. using all the mechanisms and rules explained in this chapter. After these agents had interacted with one another, /eə/ had shifted towards /ɪə/ in the speech of agents representing older speakers, as expected. Furthermore, an analysis of the sub-phonemic classes computed by GMM and NMF revealed that the contrast between the diphthongs was increasingly neutralised over the course of the simulation. More specifically, most agents organised their stored exemplars into two sub-phonemic classes after the interactions, but these classes contained a balanced mixture of words with canonical /ɪə/ and /eə/. Thus, the study by Gubian et al. (2023) shows that `soundChangeR`, in contrast to all other currently existing ABMs of sound change, can model phonological mergers by allowing for agent-specific sub-phonemic classes that are derived from the statistical properties of phonetically detailed exemplars and that are updated regularly (Coleman, 2002; Kiparsky, 2018; Scobbie, 2006; Scobbie & Stuart-Smith, 2008).

Many ABMs of sound change also critically depend on settings that determine how exemplars are produced although most of them, like `soundChangeR`, start by sampling a set of acoustic feature values from a Gaussian distribution. In the case of the ABM presented here, the agent speaker randomly chooses a word class and the Gaussian distribution is computed over all exemplars associated with that word class (and the distribution can be stabilised by means of SMOTE if necessary). In most of the other models (Blevins & Wedel, 2009; Ettlinger, 2007; Kirby, 2013; Sóskuthy, 2015; Todd et al., 2019; Wedel, 2006), however, random noise is added to the sampled values to model imprecision in reaching articulatory targets, or the effects of a phonetic bias are emulated by pushing the newly produced exemplars in the direction of the expected change. In the model by Todd et al. (2019, p. 6), for instance, produced tokens are shifted in a given direction to model "external influences such as reduction of articulatory effort" – although this rule is only applied to tokens of the pusher (and not the pushed) category to simulate a vowel push chain. This force, called bias by Todd et al. (2019), is required to trigger phonetic shifts in that model. Additionally, they add some noise to all tokens to model articulatory imprecision. The shuffling of tokens in any direction helps to maintain the within-category variance, as otherwise the Gaussian sampling might lead to a

return-to-the-mean effect. Another example is provided by Sóskuthy (2015) who implemented a bias that acts as an attractor for all tokens, but more strongly so for tokens that are acoustically far away from the bias location than for those that are closer. In theory this should model cases like /u/-fronting in many English varieties in which the vowel is affected differently by the fronting bias depending on its phonetic context (i.e. coronal or palatal consonants vs. other consonants; Harrington et al., 2008). However, given that the bias in Sóskuthy (2015) is a logistic function with the attractor located at the function's inflection point, tokens of /u/ that are overly fronted are retracted towards the bias location just like overly retracted tokens of /u/ are pulled to the front. In the ABMs by Blevins and Wedel (2009) and Wedel (2006), the addition of noise to the produced token is used as a source of variation, though they also implement measures to prevent categories from broadening inexorably – a concern that is mediated by the absolute memorisation criterion in `soundChangeR`. While the addition of random noise during the production procedure could indeed counteract the effect of decreasing within-category variance caused by Gaussian sampling in `soundChangeR`, the absolute memorisation criterion likely exerts an even stronger influence in this regard than the Gaussian sampling. That is, if a token is rendered too atypical through the addition of random noise, it will be rejected by the Mahalanobis distance threshold in our ABM, which can lead to reduction of within-category variance over time as shown in section 3.3.2.1. In these and other computational models (e.g. Ettlinger, 2007; Kirby, 2013) it remains unclear whether the added noise models articulatory imprecision or indeed phonetic bias faithfully and (as with many settings in such computational models) what amount of noise could be considered realistic based on experimental evidence.

A computational model must abstract from reality and may therefore simplify real-world processes for the benefit of achieving generalisability. However, as sound change is a multi-factorial process, there certainly are conceivable extensions to the ABM presented in this chapter. The first concerns the parameterisation and transmission of data. Although technically the ABM accepts any numeric continuous values as acoustic parameters, it has been an aim of the computational implementation of the IP model to use parameters

as input that reflect dynamic properties of the speech signal, such as DCT coefficients or PC scores (Harrington & Schiel, 2017). While it seems reasonable to store exemplars in some form of acoustic parameterisation in the agents' memories, it is questionable whether these should also be transmitted. A possible alternative would model the process of articulation more closely, i.e. during production, the sampled parameters are converted into speech signals (by means of inverse DCT or, in case of PC scores, the process explained in section 2.2.1.3). The resulting signal is then transmitted to the listener who parameterises it again and goes through the perception process. This is especially useful when the chosen parameterisation is data-specific (e.g., FPCA) because each agent could develop their own internal language model over time based on their language experience (recall Figure 2.8).

The second extension that might be considered for future versions of the agent-based model presented here is the implementation of activation levels. In exemplar theory, the production and perception of exemplars is influenced by their activation level which is linked to both the frequency of the word in which the exemplar was uttered as well as the time that has passed since its memorisation (Pierrehumbert, 2001, 2002). That is, exemplars associated with a high-frequency word and those stored more recently have higher resting activation levels. The proposal by Pierrehumbert (2002) about exemplar-based speech production states that tokens are produced by computing the average acoustic values over a few stored exemplars while giving recently stored exemplars more weight than older ones. The classification of exemplars in perception is similarly impacted by time-decaying activation levels according to Pierrehumbert (2001), i.e. the perceived token is assigned the label of the phonological class whose exemplars, weighted by activation strength, are closest to the token. This account is supported by findings from phonetic imitation experiments which might hint to a stronger activation of more recent as compared to older exemplars (e.g. Pardo, 2006) and has parallels to spreading-activation theories in lemma selection and retrieval (Levelt, 2001; Roelofs, 1992) as well as in spoken word recognition (Luce & Pisoni, 1998). Time-decaying activation levels have been implemented in a number of computational models, exerting their impact either only in speech production (Wedel,

2006) or in both production and perception (Ettlinger, 2007) or they are only used for the purpose of memory management (Kirby, 2013). In earlier versions of `soundChangeR`, the concept of time decay was implemented in the form of a forgetting strategy, i.e. the user could choose to have the agents forget the oldest exemplar (very similarly to Stanford and Kenny, 2013). Especially in small datasets, this strategy led to unforeseen random walks of the exemplar clouds through the acoustic space, which is why the agents in `soundChangeR` do not assign time stamps to exemplars and instead remove randomly chosen exemplars from memory (although it must be noted that this forgetting strategy was a rougher implementation of time decay than the gradually decreasing activation levels). Conceptually, the time decay approach also neglects recent findings on accent reversal in elderly individuals who return to using characteristics of their childhood dialect (Harrington & Reubold, 2021; Reubold & Harrington, 2015). That is, these individuals either re-activate or have never de-activated the oldest exemplars in their memory, otherwise they could not accurately produce their childhood accent in old age.

So instead of extending `soundChangeR` by time-decaying activation levels, activation levels based on lexical frequency in speech production and perception could be implemented, a proposal that has only been realised and tested by Todd et al. (2019) despite the numerous studies investigating effects of lexical frequency on sound change (e.g. Bybee, 2002, 2015; Clark and Trousdale, 2009; Hay et al., 2015; Hooper, 1976; Lin et al., 2014; Phillips, 1984; Tamminga, 2013; for an overview of frequency effects on aspects of language and memory in general see Divjak, 2019). According to Pierrehumbert (2001), exemplars associated with high-frequency words should have higher activation levels than those associated with low-frequency words. So, given that high-frequency words in such a model are produced more often, they are more often subjected to phonetic biases which eventually turn into permanent change. That is why exemplar theory predicts that high-frequency words should change faster than low-frequency words (Hay & Foulkes, 2016). If frequency-based activation levels were incorporated into `soundChangeR`, the choice of a word type in speech production would no longer be random, but the probability of a word type being chosen would depend on its activation level. Hence, high-frequency

words would be produced more often than low-frequency words. If frequency-based activation levels do not influence speech perception (e.g. because speech perception is first and foremost a phoneme-based process in `soundChangeR`), high-frequency words will be represented by more exemplars in the agents' memories than low-frequency words after some interactions – that is, if the agents have not already been initialised with a number of exemplars per word type that reflects their lexical frequency. In any case, this imbalance in the exemplar storage would result in a reduced impact of high-frequency exemplars on the shape, size, and orientation of the exemplar clouds (cf. Figure 8 in Todd et al., 2019, p. 11), thereby rendering high-frequency words less susceptible to change than low-frequency words. While this is contrary to the predictions of exemplar theory, there is some evidence in support of faster rates of change in low-frequency than high-frequency words (Hay et al., 2015; Phillips, 1984). It should also be noted that the usage of activation levels only in speech production puts focus on the speaker as a conduit for sound change, while the IP model follows a listener-based approach. It is somewhat more difficult to foresee the impact of frequency-based activation strength when integrated into the speech perception mechanisms of `soundChangeR`, since it is not word or phoneme recognition that is modelled (Clopper et al., 2010; Connine, 2004; Connine et al., 1993; Dahan et al., 2001; Forster & Chambers, 1973; Jescheniak & Levelt, 1994; Luce & Pisoni, 1998; Vitevitch & Luce, 1999), but rather probabilistic memorisation of phonetic details, categorisation behaviour, and statistical learning. So, whereas previously a token t would be accepted as an exemplar of phoneme P1 if its probability of belonging to that class was higher than that of belonging to alternative phoneme P2 – $P(t|P1) > P(t|P2)$ – and if its Mahalanobis distance to P1 was below a given threshold, both of these metrics could be weighted by the activation levels of the stored exemplars. Consequently, exemplars of high-frequency words would have a higher probability of being memorised than those of low-frequency words, even if they are ambiguous or atypical.

Another possible extension concerns generational changes in the agent population, a mechanism that has previously been implemented by others (Kirby, 2014b; Kirby & Sonderegger, 2013; Stanford & Kenny, 2013). The reason to

consider this extension is that there is evidence that shows that children can be drivers of sound change, particularly sound changes that are essentially non-social (Nardy et al., 2014; Nielsen, 2014; Roberts, 1997; Trudgill, 2004, 2008b). Trudgill has repeatedly underlined the critical role of children in the development of new dialects or languages in contact-scenarios such as colonialisation in which the groups of adults in contact speak diverse languages, but the first- and second-generation children establish new speaking norms (Kerswill & Trudgill, 2005; Trudgill, 2004, 2008b). This is because children start their language acquisition with a clean slate, collecting language experience mainly from interactions with their peers (Nardy et al., 2014), and adapting quickly and seamlessly to new phonetic variants (Nielsen, 2014; but see Flege et al., 2006), while "[f]or adults, incomplete accommodation and imperfect language learning and dialect learning are the norm" (Trudgill, 2004, p. 28). In the realm of sound changes, a new generation of speakers represents a lack of continuity (Kerswill & Trudgill, 2005), i.e. the stable language state of the parent generation can be restructured phonologically by children or an ongoing phonetic change can be accelerated (Roberts & Labov, 1995; Smith et al., 2019) or, in Labovian terms, incremented (Labov, 2007). Labov's model of transmission, incrementation, and diffusion was tested by Stanford and Kenny (2013) using agent-based simulations including generational changes. The agents in this model lived for a total of 25 time-periods (and were part of 1000 interactions per time-period) and were considered children during the first five time-periods. Children in this model always started with an empty exemplar storage and collected exemplars over the course of the interactions. The total number of agents was kept stable by balancing the rates of birth and death. When this ABM was run on artificial data that emulated the Northern City Shift in Chicago and St. Louis, it was found, *inter alia*, that children advanced ongoing changes in their parent generation and thereby contributed significantly to the progression of the sound change. Presumably, using generations like this would accelerate phonetic and phonological changes in `soundChangeR`, too. This is because the parent-generation, while slowly undergoing a change, also retains many exemplars of an older speaking style, which are not transmitted to their children. While agent generations can help to increment a sound

change, it is unlikely that they can trigger them, which would render heterogeneous agent groups (e.g. according to the apparent-time paradigm, Bailey et al., 1991) superfluous. That is, using children and parents is not the same as initialising agents with production data from older and younger real speakers. If data from young speakers is used to initialise agents in `soundChangeR`, it is because they represent speakers who have largely completed the sound change under investigation and act as attractors for speakers who have yet to undergo the change. Omitting these innovative speakers from the ABM would probably result in phonetic and phonological stability amongst the homogeneous agents, even if generational changes were implemented into the model.

A question that has not been addressed in this chapter, but also has been neglected by authors of other computational models of sound change, is when simulations should be stopped. In `soundChangeR`, the user is given the option (and responsibility) of choosing the number of interactions to be executed. However, one cannot be sure which number of interactions will be enough or exactly right to complete the changes, given the number of agents, number of exemplars per word per agent, and the initial acoustic conditions of the input data. That is, the necessary amount of interactions increases with the amount of agents, amount of speech data, as well as the amount of change left to be accomplished by the simulation. Moreover, there is no established or universally applicable metric that determines when simulated changes are actually finished. Visually, one can determine, for instance, the point at which the centroids of phoneme categories do not change anymore acoustically either across the population or within all individual agents, especially when the phonological categories are fixed (as done also by e.g. Harrington et al., 2018; Sóskuthy, 2015; Stanford and Kenny, 2013; Stevens et al., 2019). This sort of impressionistic evaluation becomes much more difficult when the flexible phonology is applied, i.e. when the sub-phonemes are agent-specific and can consist of several acoustic clusters. Using the flexible phonology module also entails another challenge: Which criterion should be used to identify a complete phonological change as created by GMM and NMF? In our experimentation with the flexible phonology module, we have found that the algorithms tend to lead to a slowly progressing fragmentation of

the phonological level after there are no more significant acoustic changes across the population, at which point we opted to stop the simulation. A further complication lies in the unclear relation between real and simulated time. Most sound changes progress over the course of several generations of speakers, i.e. a few decades or even centuries (Salmons et al., 2012), and it might be impossible to determine how many simulated interactions are the equivalent of this time period. For example, while simulating vocalic shifts in an isolated population of Antarctic winterers, Harrington, Gubian et al. (2019, p. 3331) found that one of the simulated changes extended beyond the empirically observed change, and concluded: "The discrepancy between the actual and computationally modeled changes in Antarctica could have come about because there is no predictable link between the actual time spent in Antarctica and the number of interactions in the model". Establishing a stopping criterion for simulations of phonetic and phonological changes in agent-based models will be an interesting task for the further development of `soundChangeR`.

Finally and perhaps most importantly for future research, the ABM presented here also crucially differs from other computational models of sound change in that the code to the model is openly accessible and easily expandable. We are following an open science policy because it is important that the scientific community can check and try to replicate previous results (Garellek et al., 2020; Laurinavichyute et al., 2022; Roettger et al., 2019; Winter, 2020). A consequence of making the software available is that the risk of publishing erroneous results is reduced while at the same time constructive discussions about the model's implementation are invited. Since the code is versioned with git, all adaptations of the code since 2018 (when the code was first uploaded to GitHub) are tracked and justified by means of commit messages (see https://github.com/IPS-LMU/soundChangeR). Moreover, the ABM is provided as an R package, i.e. a programming language that is commonly used in speech science, and the software itself is fully documented (see Appendix B.3) which enables researchers from all around the world to apply the model themselves without having to rely on the developers. Moreover, the agent-based model was programmed in such a way that new settings and

mechanisms can be implemented easily. The modular code will hopefully invite others to expand the ABM as needed to test new hypotheses.

4 | Modelling the Change from Pre- to Post-Aspiration in Andalusian Spanish

Abstract

This study is concerned with simulating the change from pre- to post-aspiration in /st/ clusters in Andalusian Spanish. Using the agent-based model `soundChangeR`, it was tested whether this metathesis of aspiration results in the phonologisation of post-aspirated (as compared to unaspirated) voiceless plosives. Therefor, the population of agents was initialised with exemplars of /st/ and /t/ produced by speakers who had either already undergone the sound change and thus post-aspirated the cluster or who had not yet adopted the new variant. In the simulation, the agents exchanged exemplars of /st/ and /t/ in a perception-production loop and two machine-learning algorithms were used to compute the agent-specific sub-phonemic classes that linked the exemplars to their word types. As expected, the agents did not change their acoustic representation of /t/ and the phonological separation between /st/ and /t/ was established by the end of the simulation. However, younger and older agents converged towards a common acoustic representation of /st/ that included aspiration phases on both sides of the closure, so the interactions between the agents did not lead to the expected reduction of pre-aspiration. It is discussed how the mechanisms of the ABM contributed to these results and how they could be expanded to adequately model the emergence of post-aspirated voiceless plosives in Andalusian Spanish.

4.1 Introduction

In the variety of Spanish spoken in Andalusia, /s/ is glottalised or lenited in many positions within the word so that, when it occurs before a voiceless plosive, that plosive is produced with pre-aspiration, e.g. *esto* /esto/ [ehto] (engl. this). In the past years, multiple studies have established that there is a sound change in progress by which the pre-aspirated voiceless plosives /sp, st, sk/ are becoming post-aspirated (Gilbert, 2022, 2023a; Parrell, 2012; Ruch, 2013, 2018; Ruch & Harrington, 2014; Torreira, 2007, 2012; Villena-Ponsoda, 2008). This sound change was hypothesised to be the consequence of a realignment of articulatory gestures, i.e. when the oral closure is formed in phase with the opening of the glottis, but released before the vocal folds start swinging again, post-aspiration is produced, whereas an anti-phase timing of the closure leads to pre-aspiration (Parrell, 2012; cf. Figure 2.1). The first study to test this hypothesis adequately by means of analysing time-varying acoustic signals that represented the glottal and oral gestures involved in producing voiceless plosives with and without aspiration (Browman & Goldstein, 1986, 1989) was conducted by Cronenberg et al. (2020, see chapter 2). They established that there was a trading relationship between pre- and post-aspiration in /sp, st, sk/ clusters produced by a population of Andalusian speakers that was heterogeneous in terms of age and regional origin. That is, while older and East Andalusian speakers tended to produce the clusters with pre-aspiration, younger and West Andalusian speakers used the more progressive post-aspirated variant. Both pre- and post-aspiration, however, were predictable from the timing of the closure with respect to the voiceless interval in sequences of intervocalic /s/ plus voiceless plosive.

The metathesis of aspiration in Andalusian Spanish can also be considered a case of possible phonologisation, similarly to sound changes like umlaut, vowel nasalisation (Beddor, 2009; Carignan et al., 2019; J. J. Ohala & Ohala, 1993; Zellou & Tamminga, 2014), or tonogenesis (Coetzee et al., 2018; Kang and Han, 2013; Kirby, 2014a; for an overview, see Kingston, 2011). Phonologisation is defined as "the process by which intrinsic phonetic variation gives rise to extrinsic phonological encoding" (Kirby, 2013, p. 1) and often involves the

loss of the conditioning environment while maintaining the coarticulatory effect. With respect to umlaut, for instance, the phonetic precursor is the same as in /u/-fronting (Alderton, 2020b; Fridland, 2008; Harrington et al., 2008), i.e. a high front vowel has an anticipatory coarticulatory effect on high back vowels, such that e.g. /u/ > [y]. If at some point the high front vowel disappears as the conditioning environment, a phonological contrast between /u/ and /y/ can become part of the language system, e.g. Old High German *fotiz* > German *Füße* /fysə/ (engl. feet) vs. *Fuß* /fuːs/ (engl. foot) (Kiparsky, 2015; Twaddell, 1938). Before the metathesis in Andalusian Spanish began, the contrast between /sC/ and /C/ (where C is a voiceless plosive) was cued mainly by pre-aspiration, but especially in fast speech, there was an articulatory and perceptual bias towards post-aspiration: the anti-phase timing of the closure relative to the glottal opening gesture is less stable than in-phase timing (Kelso, 1984; Oliveira & Marin, 2005; Tuller & Kelso, 1989; Wimmers et al., 1992) and post-aspiration is perceptually more salient than pre-aspiration as it is associated with higher amounts of spectral energy (Bladon, 1986; J. J. Ohala, 1990; Ruch, 2018; but see Gilbert, 2023b). That is, post-aspiration came to serve as a secondary cue to distinguish e.g. /st/ as in *pasta* (engl. pasta) from /t/ as in *pata* (engl. paw). This has been shown to be true even in varieties of Spanish such as Argentinian Spanish in which there is no ongoing change by which pre-aspiration gives way to post-aspiration (Ruch & Harrington, 2014). When listeners perceive an acoustic feature (e.g. nasalisation, vowel raising, aspiration) that serves as an indicator to the underlying form, but without being able to differentiate between the coarticulatory source and its effect, the primary and secondary cues are in a trading relationship (Fitch et al., 1979; Haggard et al., 1981; Repp, 1982) such as the one reported by Cronenberg et al. (2020) for Andalusian Spanish. While it has been shown that the process of phonologisation might be preceded by a trade-off between cues (Beddor, 2009; Carignan et al., 2021; Greca et al., 2022; Kuang & Cui, 2018), it remains obscure why the formerly primary cue should vanish given that phonological contrasts are always redundantly cued (Francis et al., 2000; Holt & Lotto, 2006; Lisker, 1986; Schertz & Clare, 2020) and which mechanisms of speech production and

perception support the enhancement of the formerly secondary cue beyond the level of coarticulation (Kirby, 2013).

Agent-based models (ABMs) can be helpful in investigating possible factors that might facilitate or prevent the phonologisation of one cue and the loss of another. In general, ABMs of sound change have been used to show how social networks influence the diffusion of phonetic or more abstract categories through a speech community (Fagyal et al., 2010; Lev-Ari, 2018; Pierrehumbert et al., 2014), how sound change might be avoided if it would result in widespread cases of homophony (Blevins & Wedel, 2009; Wedel, 2006), how Labov's notions of transmission, diffusion, and incrementation (Labov, 2007) are related to communication density in vowel chain shifts (Stanford & Kenny, 2013), how the interplay of biases and functional load can lead to different outcomes of phonetic shifts (Sóskuthy, 2015), how an exemplar-based production-perception feedback loop can reinforce phonetic biases and thus result in sound change (Harrington, Gubian et al., 2019; Harrington & Schiel, 2017), and how lexical frequency impacts the rate of change in different language scenarios (Todd et al., 2019). Of special interest to the study at hand is the series of agent-based models by Kirby (2013, 2014b) which aimed to explore how phonologisation is affected by articulatory biases, probabilistic enhancement of cues, and compensation for coarticulation (also see Kirby and Sonderegger, 2013, 2015). In these models, the agents operated in an exemplar-based production-perception loop, similarly to the ABM `soundChangeR` (Cronenberg et al., 2022) which was used here. In contrast, however, produced tokens could be pushed in a direction that increases (enhancement) or decreases (bias) contrast precision on one cue dimension, thereby influencing which cues (i.e. acoustic features) are helpful to the agent listener in recognising the intended sound. Both the likelihood and degree of enhancement of a cue in Kirby (2013) depended on its informativeness (among other measures, e.g. functional load of the contrast), i.e. on how well it separated one phonetic category from another (Schertz & Clare, 2020; Toscano & McMurray, 2010, 2012). This model was successful in simulating the phonologisation of fundamental frequency and progressive loss of voice onset time as a cue to the distinction between lenis and aspirated stops in Seoul Korean (Kang, 2014; Kang & Han, 2013;

Kirby, 2013). A more abrupt switch in attention to cues was modelled by Kirby (2014b) by allowing agent listeners to use only a subset of cue dimensions in order to categorise a perceived exemplar. More specifically, a cue was no longer considered in perception if the mean of the cue's acoustic distribution fell below a given threshold which was a crucial mechanism in modelling tonogenesis in Phnom Penh Khmer (Kirby, 2014a).

The aim of this chapter is to model the sound change from pre- to post-aspiration in Andalusian Spanish analysed in chapter 2 by means of the agent-based model described in chapter 3. In order to do so, the ABM needs input data that mirrors both the change and the distinction between aspirated and unaspirated voiceless plosives, given that this contrast does not vanish as a result of the change, i.e. post-aspiration is not generalised to phonologically unaspirated plosives. Therefore, it was decided to compose a new dataset of Andalusian Spanish that consists of productions of isolated words with /st/ (e.g. *hasta*, engl. until) and /t/ (e.g. *ata*, engl. she/he tied) from the same speakers that provided the data in chapter 2. The study at hand focuses on the alveolar plosives because of a lack of data for singleton /p/ and /k/. In the first part of this chapter, the newly composed dataset is analysed in the same way as in chapter 2, i.e. using FPCA on the time-varying probability of voicing and high-frequency energy, so as to establish dimensions of variation that are relevant to the change from pre- to post-aspiration, but also differentiate between aspirated and unaspirated voiceless plosives. In the second part, the scores extracted from FPCA are used as input to a simulation with `soundChangeR`, i.e. the ABM presented in chapter 3. The agents represent the 48 speakers of Andalusian Spanish and they exchange tokens of /st/ and /t/ in their production-perception feedback loop (cf. section 3.3.2). While this computational model does not allow for a reduction in (acoustic) cue dimensions as in Kirby (2014b), it is equipped with algorithms that regularly restructure the agents' sub-phonemic classes in response to recently memorised exemplars (cf. section 3.3.1). Thus, it is possible that the more conservative agents create different sub-phonemes for unaspirated, pre-, and post-aspirated plosives, respectively, upon encountering the strongly post-aspirated variants of more innovative agents. If the acoustic bias towards post-aspiration is captured by

the FPCA parameterisation, the innovative agents will be less prone to accepting the conservatives' pre-aspirated exemplars so that they will act as attractors for the conservative agents rather than vice versa. Therefore, it is expected that the interactions between the agents lead to the decrease of pre-aspiration and the emergence of a contrast between post-aspirated and unaspirated /t/.

4.2 Aspiration in /st/ and /t/

The cognitive-computational architecture of speech processing introduced by Cronenberg et al. (2020) proposed that a transformation such as FPCA can be applied to a cloud of memorised exemplars (Goldinger & Azuma, 2004; Pierrehumbert, 2001) to gain both phonological and distributional knowledge. Phonological knowledge entails systematic, usually population-level articulatory or acoustic patterns that differentiate one class of sounds from another. Within a phonological class, the exemplars are distributed in a multi-dimensional acoustic space which can provide information about the socio-linguistic background of the speakers who produced the exemplars. With regard to Andalusian Spanish, Cronenberg et al. (2020) found that the multi-dimensional space resulting from FPCA was systematically structured such that exemplars produced by older and East Andalusian speakers were typically located in a different part of the space than those of younger and West Andalusian speakers. However, given their limited dataset, it could only be speculated that FPCA (or a similar transformation) can differentiate between /sp, st, sk/ on the one hand and /p, t, k/ on the other given that the former are produced with aspiration anywhere surrounding the closure whereas the latter are not. This section is concerned with extracting this kind of phonological knowledge from a dataset of Andalusian Spanish that is composed of a subset of the /st/ tokens presented in chapter 2 as well as additional tokens of /t/. The singleton voiceless plosives are described as phonologically unaspirated for Spanish in general, i.e. the glottal gesture that is responsible for voicelessness and the oral gesture that forms the closure are aligned in both duration and phasing (Browman & Goldstein, 1986, 1992).

Since the method is essentially the same as in chapter 2 and the data also partially overlaps, it is expected that the results of the functional data analysis should capture the trade-off between pre- and post-aspiration in /st/ clusters that has been established previously. Additionally, we expect to find a dimension of variation in the shapes of the acoustic signals that describes the differences between aspirated clusters and unaspirated singletons.

4.2.1 Method

4.2.1.1 Participants and Material

The participants were the same as those described in chapter 2. There were 48 speakers, equally distributed across two age groups (younger and older) and two regional groups (East and West Andalusian). The younger speakers were between 20 and 36 years old (mean: 26.1 years), the older speakers were between 55 and 78 years old (mean: 66.8 years). All speakers had lived in their home towns of Seville (West Andalusia) or Granada (East Andalusia) for their whole life or at least for the last 20 years prior to the recording. The recording setup was as explained in Ruch (2013). In brief, the participants were wearing a headset microphone and the material was presented to them via SpeechRecorder (Draxler & Jänsch, 2004) on a computer screen. They were asked to speak in their natural dialect as if they were talking to a friend.

For the present study, 19 words containing /st/[9] and 9 words containing singleton /t/ were selected from the complete corpus which consists of interviews, a read text, and a list of 180 isolated words. The target words for this study (see Appendix C.1) were produced in isolation by the speakers and repeated three times, resulting in a total of 48 (speakers) × 28 (words) × 3 (repetitions) = 4032 tokens. 36 tokens were excluded because they were misread, leaving 3996 tokens for the analysis.

[9] In Cronenberg et al. (2020), a total of 20 words containing /st/ were analysed. The same word types were used here, with the exception of *pasta*. This is because *pasta* was produced in a short sentence instead of in isolation by most speakers.

4.2.1.2 Data Processing

The data was processed entirely automatically without the need for manual segmentations, alignments, or corrections. The first step was to transform the data into an EMU database (Winkelmann, 2017) and automatically segment it using a grapheme-to-phoneme converter and MAUS from the BAS Webservices (Schiel et al., 2011). In a second step, the high-frequency energy (HF) signal and voicing probability (VP) were computed as described in section 2.2.1.2 (Cronenberg et al., 2020). These two acoustic signals are used as proxies for the articulatory gestures needed to form a voiceless plosive with and without aspiration phases surrounding the closure (Goldstein & Browman, 1986): VP represents the glottal gesture, HF the oral gesture.

The alignment of HF and VP can represent closures, vowels, and aspiration phases in the speech signal (Cronenberg et al., 2020). Closures are marked by voicelessness and very low energy, i.e. both HF and VP must be low in this case. Vowels are characteristically voiced and have a lot of energy also in the higher frequencies, so a vocalic segment is represented by high HF and VP signals. In contrast to both closures and vowels, aspiration phases are typically voiceless and, given the glottal source of friction, have a lot of energy in high frequency bands. This means that aspiration is present in the speech signal when VP is low and HF is high (see Figure 2.3 for examples of a pre- and a post-aspirated token of the word *despide*). Therefore, we take the positive area between HF and VP (as computed using formulae (2.2)) as a measure for aspiration strength.

There were two deviations from the data processing procedure in Cronenberg et al. (2020) which used a database where the closure in /sC/ clusters had been segmented manually by Hanna Ruch for her dissertation (Ruch, 2013). For the present study, no prior manual segmentations were used. Thus, the first deviation from Cronenberg et al. (2020) was that the amplitude normalisation of HF was achieved by subtracting from the HF signals the lowest energy value during the production of the automatically segmented plosive, and not of the closure. The second deviation affected the sequence of interest in /V_1(s)tV_2/ which extends between the last (resp. first) point in time where VP reached

a local maximum during the production of /V_1/ (resp. /V_2/). This approach relied on the automatic segmentation by MAUS instead of searching for the VP peaks going backwards (resp. forwards) in time from the manually set boundaries of the closure. Both of these deviations are not expected to have a notable impact on the results.

As a prerequisite for the functional analysis, the pairs of HF and VP signals were linearly time normalised between the beginning and end of the sequence of interest. Additionally, HF was re-scaled to values between approx. 0 and 1 by dividing each HF signal by the third quartile over all HF data points in order to avoid that any variation in HF would overshadow variations in VP.

4.2.1.3 Functional Data Analysis

The 3996 pairs of HF and VP signals were given as input to Functional Principal Components Analysis (FPCA; Gubian et al., 2015; J. O. Ramsay and Silverman, 2010) in order to identify the main modes of variation in the signal shapes. The FPCA was executed as described in section 2.2.1.3, but using a more recent version of the R package `fda` (version 6.0.3, J. Ramsay et al., 2021). FPCA returns three main objects as can also be seen from equations (2.1a) and (2.1b). The first is the mean HF and VP signals over all input signals, $\mu_{HF}(t)$ and $\mu_{VP}(t)$. The second is a set of K pairs of time-varying Principal Components $PCk_{HF}(t)$ and $PCk_{VP}(t)$ which capture distinct modes of variation in the signal shapes. The third are PC scores $s_{k,i}$, one for each input signal i and each PCk. These scores can be considered weights that determine how much of variation k is present in input signal i. Since FPCA decomposes each pair of input signals into a linear combination of K PCs added to the mean, the PCs and PC scores can also be used to approximately reconstruct the original signals. For this analysis, the first $K = 4$ PCs were computed. Together, these described 78.7% of the variance in the signal shapes, apportioned into 31.1%, 27.8%, 12.3%, 7.6% for PC1, PC2, PC3, and PC4 respectively.

4.2.2 Results

The results are presented in two parts: The first part is about the kinds of variation captured by the PCs, i.e. the independent but systematic variations in the shapes of the HF and VP signals that are relevant to both the sound change in /st/ clusters as well as to the distinction between /st/ and /t/. The second part of the results is a statistical analysis using Linear Mixed Effect Regressions which test whether there are correlations between the PCs and the speakers' age, regional origin, and the plosive type (aspirated vs. unaspirated).

4.2.2.1 Variation Captured by Principal Components

Figure 4.1 shows the modes of variation captured independently by PCk, where $k = 1, 2, 3, 4$. This is achieved by adding to the mean HF and VP signals $\mu_{HF}(t)$ and $\mu_{VP}(t)$ (middle column) plus or minus one standard deviation of the PCk score $\pm\sigma_{s_k}$ multiplied by the corresponding $PCk(t)$.

PC1 (top row) captured both the timing of the closure from early (left panel) to late (right panel) and the height of HF from low (left) to high (right). Additionally, PC1 encoded the duration of the voiceless interval (characterised by continuous low values of VP) from short (left) to long (right). That means that negative values of s_1 were associated with an early closure (and hence, no pre-aspiration) and very little post-aspiration, whereas positive values of s_1 were associated with a longer voiceless interval, more high-frequency energy and therefore both pre- and post-aspiration phases.

PC2 (second row) described the varying height of the HF maxima as well as the closure timing from anti- to in-phase. PC2 also captured a variation in the second half of the VP signal which results in a longer (left) or shorter voiceless interval (right). Negative values of s_2 were related to relatively low HF peaks which means that there is little possibility for aspiration to occur despite the longer voiceless interval. Positive values of s_2 on the other hand were associated with post-aspiration because of the combination of high HF peaks and the in-phase timing of the closure with the voiceless interval.

PC3 (third row) captured the compression of the HF signal with a slight parallel change in VP. Very similarly to PC3 in Cronenberg et al. (2020), this vari-

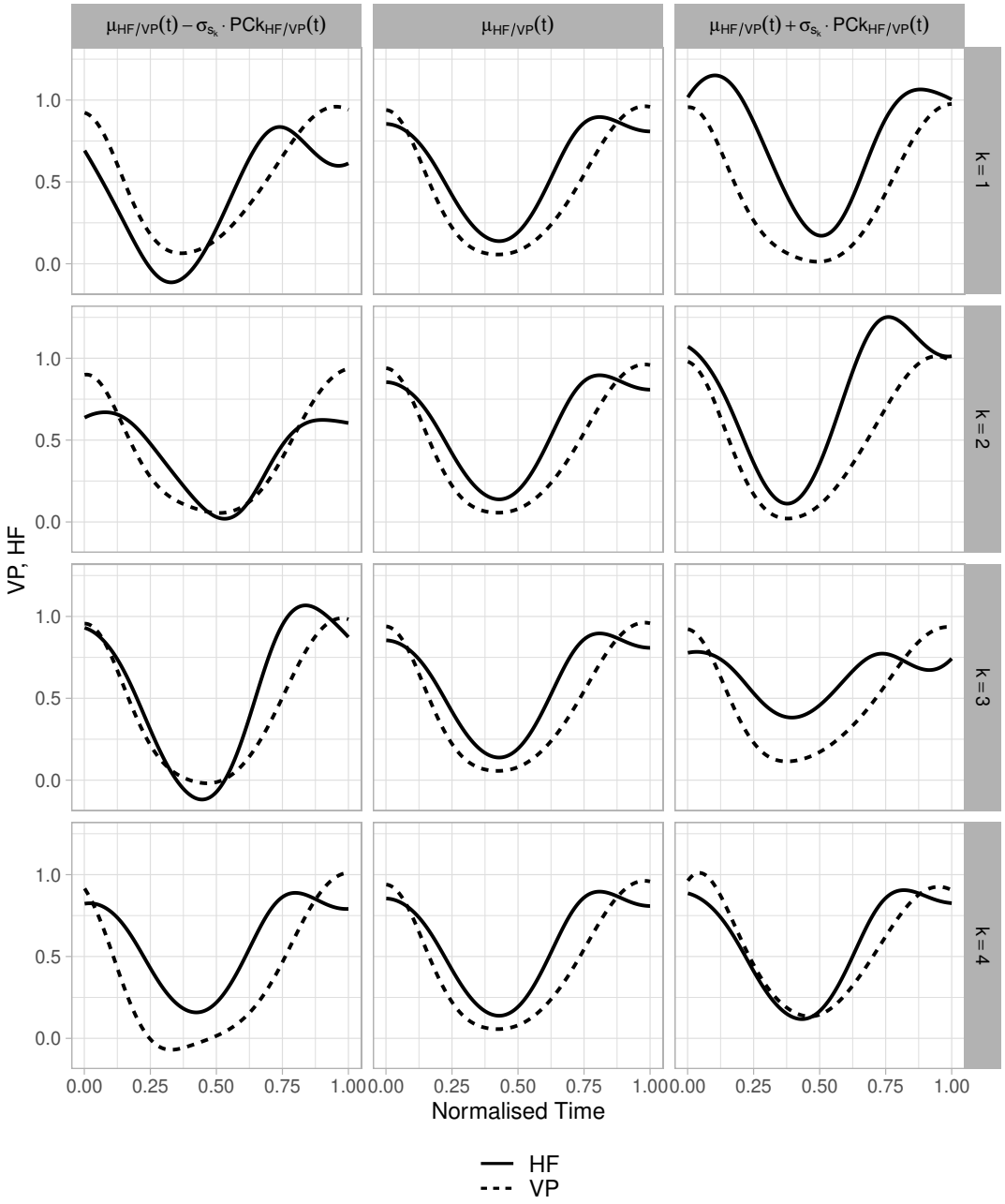


Figure 4.1: Variation expressed by the first four PCs. The middle column shows the mean signals $\mu_{HF}(t)$ and $\mu_{VP}(t)$ which were modified by subtracting from (left panel) or adding to (right panel) each mean signal the respective $PCk(t)$ signal ($k = 1, 2, 3, 4$ in rows) multiplied by the standard deviation of s_k ($\sigma_{s_k} = 0.27, 0.25, 0.17, 0.13$, for the four PCs respectively). The exact expressions are given in the column headings.

ation is unlikely to be relevant to the analysis and rather comes about because of amplitude differences between individuals or words (cf. Appendix A.5). This possibility will be further investigated with Linear Mixed Effect Regression models below.

PC4 (bottom row) described a variation in the low values of VP: Negative values of s_4 represented a long voiceless interval with VP values around zero, whereas positive s_4 values were associated with a short voiceless interval and a local minimum in the VP signal slightly above zero. In combination with HF, the variation captured by PC4 might represent aspirated (negative s_4) and unaspirated plosives (positive s_4).

4.2.2.2 Statistical Analysis

Here we focus on relations between the PC scores from FPCA and the speakers' age, regional origin, and the differences between /st/ and /t/. In line with Cronenberg et al. (2020), we expect to find a significant influence of age and region on the amount of pre- and post-aspiration in /st/, with younger and West Andalusian speakers having more post- and less pre-aspiration than older and East Andalusian speakers. We do not expect to find a similar influence of these sociophonetic variables on /t/ which should present with little to no aspiration, especially pre-aspiration.

The variation in the signals' shapes which is relevant to both the overall amount of aspiration (for the difference between /st/ and /t/) and to the sound change from pre- to post-aspiration in /st/ is spread across at least PC1, PC2, and PC4. That is why a linear mixed model was constructed for each of the four PCs with the score s_k as the dependent variable. The maximal model contained age (older vs. younger), region (East vs. West Andalusia), and plosive (/st/ vs. /t/) as well as all two-way interactions and the three-way interaction as fixed factors. Random intercepts were computed for word (28 levels) and speaker (48 levels) and random slopes were computed for age and region by word as well as for plosive by speaker. The full model in R notation is given in (4.1).

$$
\begin{aligned}
s_k \sim\ & (age + region + plosive)^3 + \\
& (age + region \mid word) + \\
& (plosive \mid speaker)
\end{aligned}
\tag{4.1}
$$

The LMERs were pruned using the `step` function from the R package `lmerTest` (version 3.1.3, Kuznetsova et al., 2017). For the LMER with s_1 as dependent variable, only the random slope for region by word was removed from the model. The same was done for the model of s_2, in addition to the removal of the three-way interaction between the fixed factors as well as the two-way interaction between age and plosive. For the model with s_3 as dependent variable, plosive was pruned as a fixed factor and all interactions between the remaining fixed factors were also removed. The random effects were left unchanged in this case. Lastly, for the LMER on s_4, all interactions between the fixed factors as well as the by-word random slopes for age and region were removed.

PC score s_1 was significantly influenced by plosive ($t[63.4] = 3.1, p < 0.01$). There were also significant interactions between age and region ($t[44.0] = 2.3, p < 0.05$) as well as between the three fixed factors ($t[44.0] = 2.7, p < 0.01$). Given the significant interactions, we computed pairwise comparisons with `emmeans` (version 1.7.4.1, Lenth, 2022). There was a significant difference between /st/ and /t/ for older East ($t[63.4] = 3.1, p < 0.01$), younger East ($t[63.2] = 2.8, p < 0.01$), and older West Andalusians ($t[63.6] = 4.3, p < 0.001$). In all of these cases, the estimated marginal mean was negative for /t/ and positive for /st/ (see Table C.3 in Appendix C.4 for the exact values), so according to the top row of Figure 4.1 there was less aspiration in /t/ than in /st/. This difference between /st/ and /t/ can also be observed in the top row of Figure 4.2 which shows boxplots of s_1 by age (colour-coded), region (x-axis), and plosive (columns). For younger West Andalusian speakers, the estimated marginal means for /st/ and /t/ were both negative and the difference between them was not significant. Referring back to Figure 4.1, this means that /st/ and /t/ are characterised by an early closure and relatively low energy when

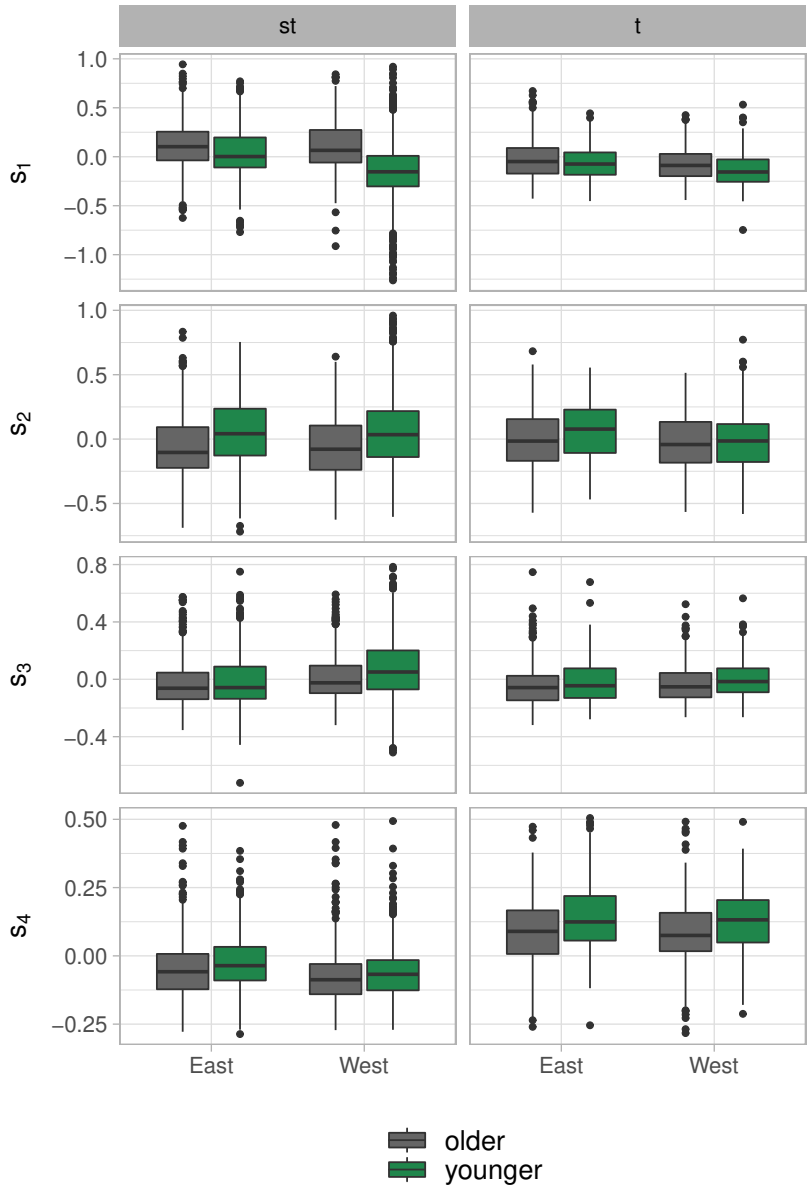


Figure 4.2: Boxplots of PC scores s_1, s_2, s_3, and s_4 (top to bottom row), separately by age group (colour-coded), region (x-axis), and /st, t/ (columns).

produced by young West Andalusians. Furthermore, there was a significant s_1 difference between young and old West Andalusian speakers producing /st/ ($t[46.7] = 4.2, p < 0.001$) which is also visible from Figure 4.2. The estimated marginal mean was negative for younger and positive for older West Andalusian speakers' /st/. Finally, there was a significant difference between younger East and West Andalusian speakers producing /st/ ($t[44] = 3.4, p < 0.01$), and an almost significant difference for /t/ ($t[44] = 1.8, p = 0.08$). The estimated marginal mean was negative for younger West, but positive for younger East Andalusian speakers' /st/. Overall, it can be summarised that younger speakers, West Andalusian speakers, and /t/ tended to be associated with lower values of s_1 (which indicates an earlier closure, shorter voiceless interval, and overall less energy) than older speakers, East Andalusian speakers, and /st/.

PC score s_2 was significantly influenced by age ($t[47.6] = 2.8, p < 0.01$) and there were significant interactions between age and plosive ($t[46.8] = 3.4, p < 0.01$) as well as between region and plosive ($t[45.1] = 2.1, p < 0.05$). The pairwise comparisons showed significant differences between older and younger speakers producing /st/ independently of the speakers' regional origin ($t[47.6] = 2.8, p < 0.01$), as shown in the left panel of the second row of Figure 4.2. The estimated marginal mean in this case was negative for older and positive for younger speakers (see Table C.4 in Appendix C.4). So according to PC2 from Figure 4.1 (middle row), younger speakers' /st/ is characterised by an in-phase closure timing and a high HF maximum in the second half of the signals, resulting in strong post-aspiration, whereas older speakers' /st/ has a long voiceless interval with a late closure leading to some pre-aspiration despite the overall low energy.

There was a significant effect of age ($t[56.4] = 2.1, p < 0.05$) and region ($t[53.0] = 2.3, p < 0.05$) on PC score s_3. Pairwise comparisons were not computed in this case, but the estimated marginal mean s_3 for older East Andalusian speakers was negative, whereas it was positive for younger West Andalusians (see Table C.5 in Appendix C.4). Recall from Figure 4.1 that PC3 captured a vertical compression of the HF signal (and also to a lesser degree of the VP signal), with positive s_3 values being associated to stronger compression. For older West and younger East Andalusian speakers, the estimated

marginal mean was close to zero (cf. third row in Figure 4.2 which shows the s_3 values also separately for /st/ and /t/ even though plosive was pruned from the LMER). As stated earlier, the variation captured by PC3 is difficult to interpret in the context of the sound change, but might be capturing systematic differences in how the re-scaling of the HF signal affected the age and regional groups.

The LMER with s_4 as dependent variable resulted in significant effects for age ($t[45.0] = 3.2, p < 0.01$), region ($t[45.0] = 2.1, p < 0.05$), and plosive ($t[38.5] = 12.0, p < 0.001$). The bottom row of Figure 4.2 shows that s_4 clearly differentiates between /st/ and /t/. In line with this, the estimated marginal means in all four speaker groups was negative for /st/, but positive for /t/ (see Table C.6 in Appendix C.4). Referring back to Figure 4.1, this means that /st/ is associated with a longer voiceless interval and overall more aspiration than /t/.

Note that those PCs that are relevant to the estimation of aspiration can complement or cancel out each other's influence. For example, PC scores s_1 and s_4 determined that /st/ (in contrast to /t/) was associated with a long voiceless interval and high energy levels, without being able to differentiate between pre- and post-aspiration. This information was provided by the results on s_2 which showed that /st/ for younger speakers was associated with post-aspiration and for older speakers with pre-aspiration. Figure 4.3 shows the VP and HF signals, reconstructed using all four PCs as well as the median PC scores for each combination of age × region × plosive in formulae (2.1). For /st/, it can be observed that older speakers have more pre-aspiration (yellow) than post-aspiration (blue), whereas younger speakers produced /st/ with more post- than pre-aspiration. From older East to younger West Andalusian speakers, the closure also shifts from anti- to in-phase articulation. The unaspirated singleton /t/ presents with some post-aspiration, but no pre-aspiration according to the right column of Figure 4.3. In comparison, however, there is less post-aspiration in /t/ than in /st/. Furthermore, the aspirated cluster is associated with a much more extensive voiceless interval than /t/ in all four speaker groups.

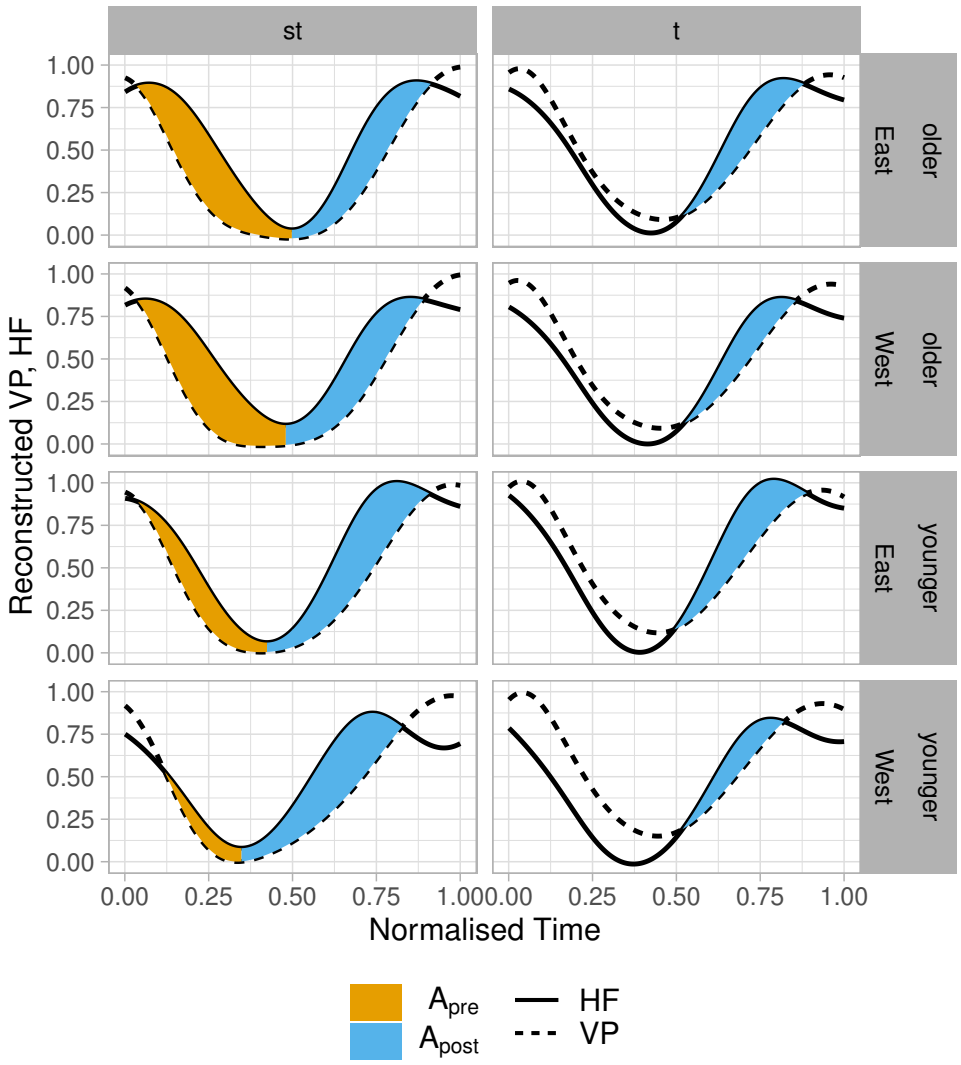


Figure 4.3: HF and VP reconstructed using formulae (2.1) in which s_k was replaced by the median s_k for each factor combination (i.e. each combination of the levels of age, region, and plosive). All four PCs were used for this reconstruction, so $k = 1, 2, 3, 4$. The median s_k values are provided in Table C.1.

4.2.3 Discussion

This section provided a functional (i.e. dynamic) analysis of a database of /st/ and /t/ produced by Andalusian Spanish speakers of two age and two regional groups. The analysis relied on acoustic signals that represented the glottal and oral articulatory gestures that are needed to produce voiceless plosives both with and without aspiration (Goldstein & Browman, 1986): the voicing probability (VP) and the high-frequency energy signal (HF). Functional Principal Components Analysis determined systematic variations in the shapes of these signals that differentiated between aspirated and unaspirated plosives as well as between pre- and post-aspirated clusters. PC2 captured how /st/ was systematically produced with pre-aspiration by older, but with post-aspiration by younger speakers. This finding supports earlier studies which found a sound change in progress from pre- towards post-aspirated /sC/ clusters in Andalusian Spanish (Cronenberg et al., 2020; Gilbert, 2022; Parrell, 2012; Ruch & Harrington, 2014; Torreira, 2007) by means of an apparent-time approach (Bailey et al., 1991). Overall, the analysis has also provided further support for modelling this change as a realignment of the closure with respect to the voiceless interval, i.e. as a shift from anti- to in-phase timing.

PC1 and PC4 mainly separated /st/ from /t/: according to this parameterisation, the cluster is characterised by a longer voiceless interval, more high-frequency energy and therefore more aspiration. This shows for the first time that the acoustic parameterisation introduced by Cronenberg et al. (2020) is also capable of describing singleton voiceless plosives that are not aspirated. While PC1 and PC4 did not differentiate between pre- and post-aspiration, they captured that /t/ was produced with a small amount of post-aspiration despite being phonologically unaspirated in Andalusian Spanish. This is not entirely unexpected given that the release of the closure is bound to let air escape from the oral cavity. Furthermore, previous studies which used more traditional duration measurements have found that the voice onset time after /t/ is around 20 ms in Andalusian Spanish (Ruch, 2013; Torreira, 2006, 2007). Nevertheless, and in line with the cognitive-computational model of sound change proposed by Cronenberg et al. (2020), the analysis has shown that FPCA

can provide phonologically functional knowledge about the different ways in which aspirated and unaspirated voiceless plosives are produced in this variety. This study has thereby extended the one in chapter 2 in an important way: it has shown that the contrast between underlying /sC/ clusters and /C/ (where /C/ is a voiceless plosive) is maintained despite the aspiration metathesis in /sC/, i.e. /C/ remains unaffected by the change towards post-aspiration.

4.3 Agent-Based Simulation

This part of the study is concerned with applying the agent-based model `soundChangeR` presented in chapter 3 to the data assembled and analysed in the previous section. The aim here is to model the change from pre- to post-aspiration in Andalusian Spanish, i.e. a sound change that entails both acoustic and phonological components (O'Neill, 2009). According to e.g. Parrell (2012), the realignment of the articulatory gestures (or, in terms of the chosen parameterisation, of the HF and VP signals; Cronenberg et al., 2020) affects the phonetic realisation of /st/ such that pre-aspiration diminishes and post-aspiration increases. Additionally, this change might lead to a phonologisation of post-aspirated voiceless stops in Andalusian Spanish which would contrast with unaspirated voiceless stops. That is, the phonological contrast between e.g. /st/ and /t/ is upheld, but is cued by post- instead of pre-aspiration.

As detailed in chapter 3, it is necessary to include data from heterogeneous speakers, some of which lie in the direction of the change. This is because there is no mechanism in `soundChangeR` that emulates a production bias by e.g. pushing the produced token in the direction of the change. Instead, it can be observed whether the initial equilibrium between two agent groups, one of which is more advanced with regard to the sound change under investigation, shifts towards the more progressive agents as a result of the interactions. In this case, it was decided to focus on interactions between age groups (i.e. older vs. younger) because age was a stronger predictor for the sound change than regional origin (also see Cronenberg et al., 2020). The singleton plosive /t/ was included as a sanity check for two reasons: first, /t/ is not expected to change

in Andalusian Spanish and second, it provides a phonological contrast to /st/ which should remain stable even after the sound change in /st/ is complete.

The hypotheses were constructed separately for the acoustic and phonological outcome of the simulation.[10] All agents should produce the cluster /st/ with more post-aspiration than pre-aspiration after than before the simulation. That is, especially those agents that represent older speakers should adopt PC score values for /st/ that resemble those of younger agents. The acoustic representation of /t/, on the other hand, should not change due to the interactions. Phonologically, it is expected that the flexible phonology module in `soundChangeR` should detect two sub-phonemic classes that correspond to the canonical underlying phonemes /st/ and /t/. There should not be any splits or mergers, i.e. the phonological separation between /st/ and /t/ should be maintained.

4.3.1 Method

4.3.1.1 Simulation Settings

The input data to the simulation were the four PC scores extracted from the acoustic data in section 4.2. In total, there were 3996 tokens distributed across 19 word types with /st/ and 9 words types with /t/ and 48 speakers. The R package `soundChangeR` was used to perform the simulation. Each of the 48 speakers was represented by a computational agent who exchanged exemplars across age group for a total of 250,000 interactions. Before the first interaction, the agents' memories were expanded by a factor of five in order to render the computation of Gaussian distributions throughout the simulation more robust. Although no phonological change (in the sense of a split or merger) was expected, the flexible phonology module was applied, i.e. sub-phonemes were agent-specific and regularly updated using the combination of GMMs

[10] At the current stage of the ABM's development, hypotheses about simulation outcomes are derived from theoretical or empirical knowledge about the direction and effects of a sound change, and not from knowledge about or experience with the mechanics of `soundChangeR`. That is, `soundChangeR` is working as intended only if it manages to produce results that align with our hypotheses; otherwise the simulation results are an indication of shortcomings of `soundChangeR` which require extensions or adjustments.

and NMF as explained in section 3.2.5. Both the absolute and relative decision criterion had to be passed in order for a new token to be memorised. After each memorisation, the agent listener removed a random exemplar of the same word class as the newly memorised token from its memory. Forgetting was blocked if, as a result, there would be less than 15 exemplars of that word in the agent listener's memory. The simulation was repeated 10 times in order to test for the robustness of the results, but the repetitions were indeed so similar that all following analyses and results are reported for only one simulation run.

4.3.1.2 Quantification of Changes

Changes that resulted from the simulation were quantified separately for the acoustic and phonological levels. Acoustic changes were analysed statistically by means of four linear mixed effect regressions with the PC scores as dependent variables. The full model is given in Eq. 4.2 in R notation and was pruned using the `step` function. The fixed effects were group (older vs. younger), simulation state (baseline vs. post-run), and plosive (/st/ vs. /t/). Random effects were included for word (28 levels) and speaker (48 levels). For the models with s_1 and s_4 as dependent variables, only the by-speaker random slope for state was removed from the models. In addition to this, the by-word random slope for state was pruned from the LMER with s_2 as dependent variable. For the model with s_3 as dependent variable, the fixed factor for age group was removed as well as the random slope for state by speaker. It must also be mentioned that post-hoc comparisons with `emmeans` could not be computed because there were approx. 50,000 data points in the dataset that resulted from the simulation. Therefore, 13,420 data points were randomly sampled from the simulation result.[11] The results of the LMERs did not differ in any meaningful way between the full and sub-sampled dataset, but since post-hoc comparisons can only be reported for the latter, we also report the LMERs for

[11] 5 samples × 48 agents × 28 words × 2 simulation states = 13,440 data points; since four agents started out with zero exemplars of the word *estuche*, the sub-sampling resulted in a dataset with 13,440 − 20 = 13,420 observations.

the sub-sampled data.

$$\begin{aligned} s_k \sim\ & (group + state + plosive)^3 + \\ & (group + state \mid word) + \\ & (state + plosive \mid speaker) \end{aligned} \tag{4.2}$$

Phonological changes were quantified by means of two measures. First, the agreement A between the sub-phonemic classes and the canonical phonemes /st/ and /t/ was computed as follows (formula taken from Gubian et al., 2023):

$$A = \frac{1}{N_w} \cdot \sum_p N_{\text{majority}}(p), \tag{4.3}$$

In Eq. (4.3), "N_w is the total number of word classes and $N_{\text{majority}}(p)$ is the number of unique word classes in the sub-phoneme p that belong to the canonical phoneme, which counts the largest number of unique word classes in p" (Gubian et al., 2023, p. 98). When some sub-phonemic classes contain only exemplars of /st/ and others only of /t/, $A = 1$. The chance level, i.e. the level at which sub-phonemes contained a balanced mixture of exemplars of both /st/ and /t/, was $\frac{19}{19+9} \approx 0.68$) in this case.

The second measure was the number of sub-phonemic classes. The two measures complement each other; in other words, on their own they can be misleading. For example, a high number of sub-phonemes might be misinterpreted as a split of phonological classes – but if the agreement is very high at the same time, the many sub-phonemes actually correspond nicely to canonical phonemes. On the other hand, if the agreement $A = 1$, one might think that the canonical phonemes were maintained or no merger has taken place, but A might be 1 because there is only one sub-phoneme (which obviously includes all exemplars). The same statistical models as in Gubian et al. (2023) were used to analyse whether the number of sub-phonemes N_p and the canonical agreement A differed between older and younger agents (factor *group*), as well as between the baseline and post-run (factor *state*). Formula (4.4) specifies the binomial mixed model which was computed to evaluate the canonical agreement, using the function `glmer` from the R package `lme4` (version 1.1.30,

Bates et al., 2015):

$$c(N_{\text{majority}}(p),\ N_w\ -\ N_{\text{majority}}(p))\ \sim\ group\ *\ state\ +\ (1\mid agent) \tag{4.4}$$

The dependent variable in this model was a vector that was composed of entities that were already defined for equation (4.3). The agent group and simulation state, as well as their interaction, constituted the fixed factors. A random intercept for the individual agents was computed as well (a by-agent random slope for state was removed from the model due to singularity).

Differences in the number of sub-phonemic classes between the agent groups and simulation states were computed by means of a Poisson regression, using the function `glm` with the following formula:

$$N_p\ -1\ \sim\ group\ *\ state \tag{4.5}$$

The number of sub-phonemes N_p minus one was the dependent variable in this case. This was because the "Poisson distribution models counts from zero to infinity, while the number of sub-phonemes N_p can vary between one and the number of word classes N_w. Subtracting one from N_p matches the minimum value to the Poisson distribution" (Gubian et al., 2023, p. 99). The factors in this regression model were group and state as well as their interaction.

4.3.2 Results

Figure 4.4 shows HF and VP signals that were reconstructed using the median PC score values (see Table C.2) by age group, simulation state, and plosive. Additionally, the pre- and post-aspiration areas A_{pre} and A_{post} are colour-coded. For /st/, older agents slightly shifted their closure to an earlier point in time, resulting in less pre- and more post-aspiration (compare first and second row, left column in Figure 4.4). Contrary to the expectations, younger agents produced more pre- and less post-aspiration after as compared to before the simulation (compare third and fourth row, left column in Figure 4.4), thus adapting to the older agents' /st/. In both agent groups, the acoustic

representation of /t/ hardly changed (see right column), i.e. /t/ remained characterised by some post-aspiration according to A_{post} and no pre-aspiration.

This visual impression of the acoustic changes was tested statistically using LMERs as described in 4.3.1.2. The first PC score s_1 was significantly influenced by the agents' age group ($t[52.3] = 6.8, p < 0.001$), the plosive ($t[42.1] = 6.0, p < 0.001$), and the simulation state ($t[36.3] = 9.0, p < 0.001$). All two-way and the three-way interaction were significant as well, which is why post-hoc comparisons were computed. The resulting estimated marginal means as well as the p-values are visualised in Figure 4.5 (top row). There were significant differences between the baseline and post-run – i.e. changes as a result of the interactions between agents – for older agents' /st/, younger agents' /st/, and older agents' /t/. The estimated marginal means revealed that older and younger agents changed their s_1 values for /st/ to approximately the same degree, thus meeting in the middle on the PC dimension that captured both a shift of the closure from in- to anti-phase and changes in the overall energy level according to Figure 4.1. The s_1 value for older agents' /t/ decreased slightly over the course of the interactions (cf. the predicted s_1 for older agents in the top left panel of Figure 4.5). Moreover, there were significant differences between older and younger agents at the baseline producing /st/ and /t/, but no significant differences between the age groups after the simulation. That is, the interactions resulted in a convergence between the agents in s_1 for both types of plosives. The differences between /st/ and /t/ were significant for older agents at the baseline ($t[42.1] = 6.0, p < 0.001$) and post-run ($t[51.6] = 5.4, p < 0.001$), as well as for younger agents at post-run ($t[58.6] = 6.1, p < 0.001$). There was no significant difference between the plosives for younger speakers at the baseline which was to be expected given the same result from section 4.2.2.2 for the most progressive group of younger West Andalusian speakers.

The second PC score s_2 was significantly influenced by age group ($t[52.0] = 5.5, p < 0.001$) and state ($t[13270] = 10.6, p < 0.001$), but not plosive. All interactions between the fixed factors were significant as well and the results of the pairwise comparisons are shown in Figure 4.5 (second row). There were significant s_2 differences between older and younger agents producing /st/ at the baseline, i.e. before the interactions. Older agents' estimated marginal mean s_2

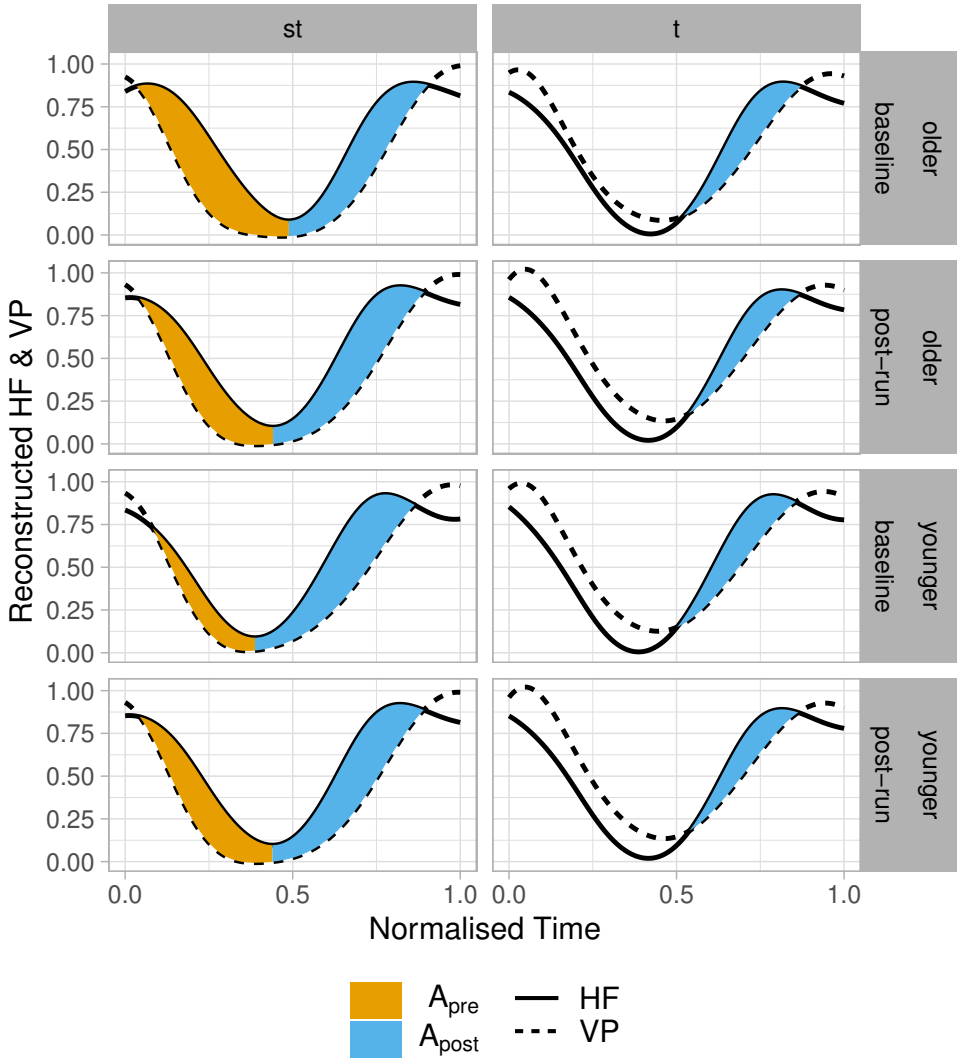


Figure 4.4: HF and VP, reconstructed using the median s_1, s_2, s_3, and s_4 from one simulation run for each factor combination of age group × plosive × simulation state in formulae (2.1). The median PC scores are provided in Table C.2.

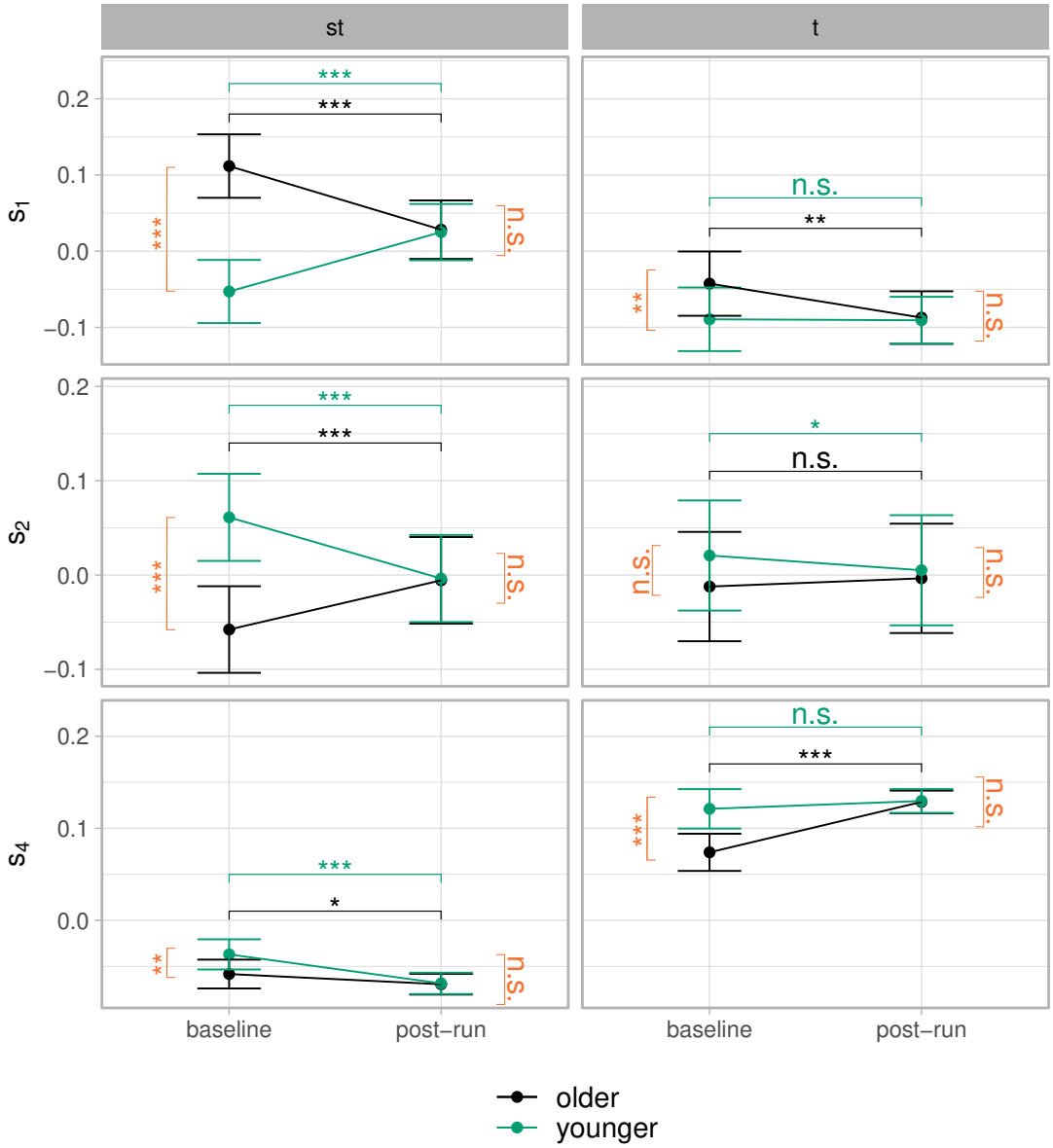


Figure 4.5: Estimated marginal means with 95% confidence bars for s_1, s_2, and s_4 (rows) for older and younger agents' (colours) /st/ and /t/ (columns) at the baseline and post-run (x-axis). The brackets with p-value indications ($p < 0.001$: ***, $p < 0.01$: **, $p < 0.05$: *, $p > 0.05$: n.s.) show statistical results from pairwise comparisons between baseline and post-run (black for older and green for younger agents), and between older and younger agents (orange). Pairwise comparisons between /st/ and /t/ and results for s_3 are reported in the text.

was negative, thus indicating an anti-phase closure in older agents' /st/ for the simulation, whereas it was positive for younger agents, indicating an in-phase timing of the closure and high HF peak in the second half of the signal (cf. Figure 4.1). This significant difference therefore differentiates between older agents' pre-aspirated and younger agents' strongly post-aspirated cluster at simulation start. There were significant changes over the course of the interactions (i.e. significant differences between the baseline and post-run) for older and younger agents' /st/ as well as for younger agents' /t/. Similarly to the changes in s_1 that resulted from the simulation, older and younger agents' converged towards a common s_2 value for /st/. Since this post-run value was very close to zero, the agents' /st/ was associated neither with strong pre- nor post-aspiration after the simulation (see second row, middle panel of Figure 4.1). This is a crucial result, given that we observed from Figure 4.4 that younger agents unexpectedly developed a more pre-aspirated /st/ due to the contact with older agents. The pairwise comparisons also showed that younger agents' /t/ had a slightly lower s_2 value after as compared to before the simulation, thereby also accommodating to the older agents (see centre right panel in Figure 4.5).

For PC score s_3, there was a significant influence ($t[26.4] = 5.9, p < 0.001$) of simulation state and an almost significant influence of plosive ($t[36.8] = 1.9, p = 0.07$). The interaction between the two fixed factors in this model was also significant ($t[26.9] = 2.3, p < 0.05$), given that there were changes in s_3 over the course of the interactions only for /st/, but not for /t/, and that the cluster was different from the singleton only at the baseline, but not post-run. More specifically, s_3 of /st/ decreased over simulation time resulting in an estimated marginal mean s_3 value for /st/ that was very similar to that of /t/. This is then why /st/ and /t/ are statistically the same in terms of s_3 after the simulation. In light of the unclear interpretation of PC3, it is difficult to state these changes' influence on the agents' speech with regard to the sound change from pre- to post-aspiration.

Finally, PC score s_4 was significantly influenced by age group ($t[55.5] = 2.9, p < 0.01$), plosive ($t[34.6] = 11.4, p < 0.001$), and simulation state ($t[32.6] = 2.4, p < 0.05$). All interactions between the fixed factors were significant, too.

The pairwise comparisons revealed significant differences between the baseline and post-run for older agents' /st/, younger agents' /st/, and older agents' /t/ (cf. Figure 4.5, bottom row). For both older and younger agents, the s_4 value of /st/ decreased slightly as a result of the interactions. Thus, after the simulation, the agents' cluster was associated with a longer voiceless interval and possibly overall more aspiration than before the simulation (cf. bottom left panel in Figure 4.1). Older agents significantly increased s_4 for /t/ due to the contact with younger agents, which means that older agents' /t/ was produced with a shorter voiceless interval and less aspiration after the interactions. Furthermore, there were significant s_4 differences between older and younger agents at the baseline producing /st/ and /t/, but not post-run, showing that the agents converged towards similar s_4 values. Lastly, there were significant s_4 differences between /st/ and /t/ for all combinations of age group and simulation state with p-values lower than 0.001. This shows that s_4 remained an important acoustic dimension for separating between /st/ and /t/ for both agent groups and throughout the simulation run.

Figure 4.6 shows the number of sub-phonemes (top row) and agreement between sub-phonemes and canonical /st, t/ (bottom row) by age group (in columns) across simulated time. Older agents started out with 2.2 sub-phonemic classes on average, while younger agents initially had about 2.4 sub-phonemes. In both agent groups, the number of sub-phonemes then dropped to approx. 1.4 to 1.5 sub-phonemes, before increasing again until it stabilised at 2 sub-phonemes.[12] According to the Poisson regression model, there were no significant differences in number of sub-phonemes between the baseline and post-run as well as between older and younger agents.

The canonical agreement steadily increased from about 0.75 to 1 in both agent groups, i.e. the sub-phonemes increasingly contained exemplars of words with either /st/ or /t/, not a mixture of both. The binomial mixed model revealed that there was a significant increase in canonical agreement from baseline to post-run ($z = 8.5, p < 0.001$), but no significant differences between

[12] Contrary to our previous experiences with the flexible phonology module, there was no increase of sub-phonemic classes after the acoustic changes had finished even when the simulation was prolonged to 500,000 interactions.

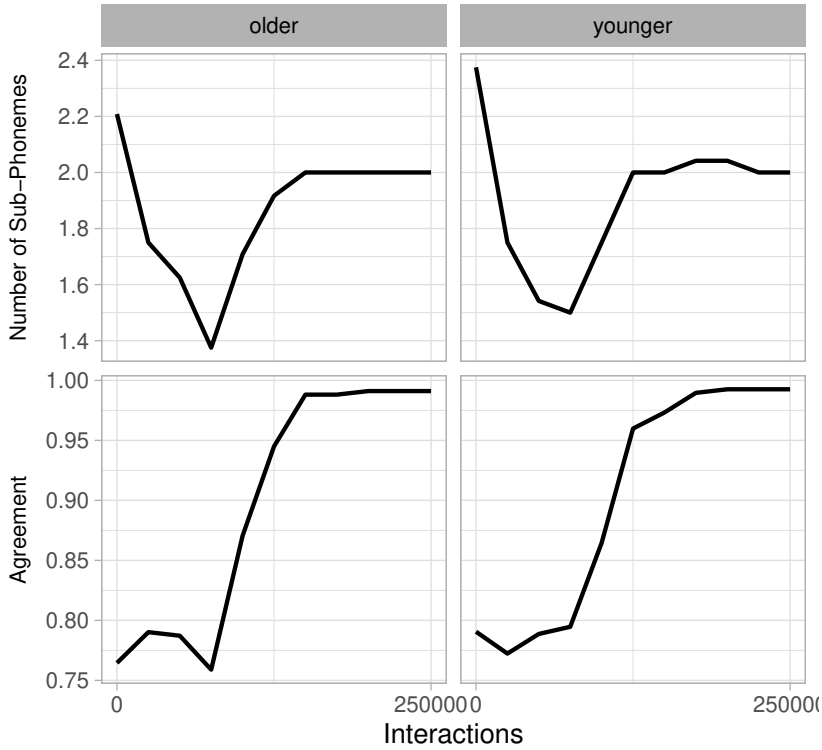


Figure 4.6: Number of sub-phonemes and agreement between sub-phonemes and canonical /st, t/ over the course of 250,000 interactions between older and younger agents.

older and younger agents at either simulation state. So while the phonological separation between /st/ and /t/ was not entirely clean in the beginning, the agents learned to organise their phonological space in such a way that the aspirated cluster and unaspirated singleton plosive became clearly distinct sub-phonemes.

4.3.3 Discussion

In summary, the results of the simulation were mixed. On the phonological level, the unsupervised machine learning algorithms adequately identified and subsequently maintained two sub-phonemic classes which, after the simulation, corresponded perfectly to the canonical phonemes /st/ and /t/ in all agents. Also in line with the expectations, there were no relevant acoustic

changes in /t/ in either agent group such that the singleton voiceless plosive remained slightly post-aspirated throughout the simulation. However, younger agents adapted their acoustic representation of /st/ at least as much to that of the older agents as vice versa, even though it was expected that older agents would adopt the younger agents' mostly post-aspirated variant of /st/. Instead, all agents produced the underlying cluster with an approximately equal amount of pre- and post-aspiration as a result of the interactions.

These two main findings, i.e. the identification of adequate sub-phonemes and the convergence in /st/, are connected to each other, but they developed at different rates. At simulation start, the flexible phonology module identified two or three sub-phonemes in each agent's exemplar storage. This slightly higher than expected number of sub-phonemes in combination with the relatively low agreement of approx. 0.75 to 0.80 (with 0.68 being the chance level and 1.00 indicating full agreement between sub-phonemes and the canonical separation between /st/ and /t/) is likely due to the four-dimensional acoustic space. In a high-dimensional space, GMMs tend to create a larger amount of acoustic clusters, such that NMF can subsequently identify more sets of clusters that pass the purity threshold and therefore form a sub-phoneme. When the agents began to interact, they started to converge especially in the s_1 and s_2 dimensions (cf. Figure 4.5) which were associated with the trade-off between pre- and post-aspiration in /st/ (see section 4.2.2). This convergence happened very quickly and was completed after 100,000 interactions. As a result, at around 75,000 interactions, there was a sudden decrease in the number of sub-phonemic classes (cf. Figure 4.6) because most agents could not devise more than one sub-phoneme in a cloud of exemplars which have no apparent structure or pattern in at least two out of four acoustic dimensions. Once this movement of agents towards each other in s_1 and s_2 had finished, s_4 began to separate /st/ from /t/ increasingly clearly in both agent groups (cf. bottom row of Figure 4.5). This allowed the flexible phonology module to once again differentiate /st/ and /t/ on a sub-phonemic level, leading to two sub-phonemes and the highest possible agreement.

The convergence between pre- and post-aspirated /st/ most likely came about because the older agents' distributions of PC1 and PC2 scores were

not skewed towards those of the younger agents (Harrington & Schiel, 2017). According to the IP model (Harrington et al., 2018), phonetic shifts result from the continuous application of a phonetic bias which, in turn, leads to directional or asymmetric phonetic distributions of sounds. In the previous chapter, it was demonstrated that a precondition for such phonetic shifts in `soundChangeR` was that one agent group's acoustic distribution of a phoneme was skewed towards that of another agent group and that the absolute memorisation criterion was applied during perception (cf. Figure 3.6 and Figure 3.10). Even though there is a phonetic bias that affects /sp, st, sk/ in Andalusian Spanish and favours post- over pre-aspiration – namely a faster speech rate (Parrell, 2012; Terrell, 1980) – this bias is apparently not reflected in the relevant PC score distributions. There are two possible solutions to this problem which could be attempted in future studies. The first is to include words with /st/ that have been produced at a faster speech rate. This should ensure that the bias towards post-aspirated /st/ is represented more strongly in the data, and subsequently also in the extracted PC scores. The other solution is to explore alternative or extended parameterisations of the data that may identify a skewed variation in the older speakers' acoustic representation of the cluster while maintaining a clear separation between /st/ and /t/.

4.4 General Discussion

There were two main findings from this study: The first was that the dynamic analysis method introduced by Cronenberg et al. (2020) can also adequately differentiate between aspirated and unaspirated voiceless plosives (i.e. between /st/ and /t/) while still capturing the trade-off between pre- and post-aspiration in /sC/ clusters. This is in line with descriptions from articulatory phonology according to which voiceless plosives with and without aspiration are the result of two gestures (glottal opening and oral closure), as well as their relative duration and alignment (Browman & Goldstein, 1986, 1989). The second was that the change from pre- to post-aspiration could not be simulated to its full extent using the ABM `soundChangeR` that is based on the IP model (Harrington et al., 2018). Here we discuss two possible reasons

for this outcome: the handling of exemplars that are acoustic outliers and the necessary components to model phonologisation computationally. Finally, we point towards potential adaptations of the agent-based model that may help to simulate sound changes that have both phonetic and phonological components, like the change from pre- towards post-aspiration in Andalusian Spanish.

4.4.1 Handling Outliers

Apart from the increase in younger agents' pre-aspiration which was adopted from older agents, another problematic result from the simulation was the decrease in younger agents' post-aspiration of /st/. More specifically, some of the more extreme s_2 values associated with strong post-aspiration (cf. Figure 4.1 and second row of Figure 4.2) were gradually removed (i.e., forgotten) by the agents and no new extreme exemplars were acquired. Three mechanisms in the ABM are responsible for that: First, with a forgetting rate of 1, the agent listener has to remove an exemplar from memory after having memorised a new one. When applied over a long enough period of simulated time (i.e., 250,000 interactions in this case), the forgetting procedure results in the removal of all originally stored exemplars – and that includes those that are acoustic outliers. Second, the production algorithm samples new exemplars from a Gaussian distribution in which extreme values occur very rarely by definition. And third, even if an outlier has been produced by an agent speaker, the absolute criterion is likely to prevent that extreme exemplar from being memorised by an agent listener. So the agent-based model is inherently biased against outliers, i.e. acoustically extreme exemplars. However, what if outliers that lie in the direction of the sound change (and, perhaps, were produced by the most innovative speakers) play an important role in the progression of said change (Labov et al., 2010; Uehara & Evans Wagner, 2017)? After all, sound change is the process of establishing an innovative variant of a sound as the new norm.

First of all, it has been shown that listeners are perceptive to the stochastic distribution of a sound and are therefore cognitively capable of detecting outliers. In an experiment in which participants listened to sounds drawn from a Gaussian distribution, Garrido et al. (2013) found that sounds from the tails

of the distribution elicited a mismatch negativity (MMN) response. This well-known effect is associated with surprise, i.e. the participants had identified some regularity in the sounds to which they were listening (statistical learning; Idemaru and Holt, 2011; Lehet and Holt, 2017; Pierrehumbert, 2003a) and clearly recognised the outliers as such (also see e.g. Cheng et al., 2010; Daikhin and Ahissar, 2012; Winkler et al., 1990). Besides their cognitive salience, outliers can also carry social meaning. For Andalusian Spanish, for example, Ruch (2018) has shown that listeners from both East and West Andalusia reliably associate strong post-aspiration in /sC/ clusters with younger and West Andalusian speakers. For diverse English varieties, it has also been shown that listeners can identify speakers' ethnicity (Purnell et al., 1999; Wong & Babel, 2017), region of origin (Clopper and Pisoni, 2004; Jacewicz et al., 2021; McCullough et al., 2019; for an overview, see Clopper and Pisoni, 2005), social class (Alderton, 2022), and make judgements of their overall personality (Alderton, 2020a) through certain acoustic cues. That is, in these cases, listeners are aware that a linguistic feature has come to be indicative of some socio-indexical characteristics of the speaker. While this feature must not necessarily be an acoustic outlier in the speaker's production, it certainly is in the listener's perception.

It may be phonetic imitation that provides the missing link between the cognitive salience and sociolinguistic categorisation of acoustic outliers on the one hand, and their contribution to sound change on the other. Several studies have emphasised that for a sound change to be initiated, there have to be innovative individuals in whose speech the effect of coarticulation on a sound is unusually pronounced (Garrett & Johnson, 2013; Stevens et al., 2019; Yu, 2021). In particular, Baker et al. (2011, p. 348) have proposed a path towards sound change in which "a hearer interprets an extreme instance of a phonetic effect as a distinct production target, an exaggeration of the normal coarticulation". The phonetically motivated and extreme variation of a sound as well as its interpretation as a new target are such rare incidents that this theory by Baker et al. (2011) adequately predicts sound changes to be rare themselves. Similarly to Baker et al. (2011), but possibly with a more explicit focus on imitation, the IP model (cf. section 3.1 and Harrington et al., 2018;

Harrington and Schiel, 2017) as well as the sociolinguistic account by Labov (1990, 2001) claim that innovative speakers may be imitated more often than their peers. If innovative, strongly coarticulated variants of a sound are very likely to be imitated even though, at the same time, they are very unlikely to occur frequently, they must be weighted more heavily when they are encoded in the listeners' memories. This is generally in line with exemplar models which propose that stored episodes of speech receive activational weights which affect their influence on both speech production and perception (K. Johnson, 1997; Pierrehumbert, 2001, 2002) as well as with models suggesting that the exemplar storage is structured to some degree in terms of meaningful socio-indexical categories (Creel & Bregman, 2011; Hay et al., 2006; K. Johnson, 2006; Kleinschmidt et al., 2018; Munson, 2011). More specifically, Sumner et al. (2014) propose that exemplars are weighted according to their social salience, whereas the exemplar's typicality plays a less important role.[13] That is, atypical and socially salient tokens (e.g. strongly post-aspirated tokens of /st/) are given a higher weight than non-salient tokens regardless of their typicality (e.g. typical: a token with both pre- and post-aspiration; atypical: a strongly pre-aspirated token) (Sumner et al., 2014).

This suggests how the agent-based model `soundChangeR` could be extended to counteract the inherent bias against outliers, given their role in sound changes. First, as proposed in chapter 3 (section 3.4), perceived exemplars would have to be given a weight before they are stored in an agent's memory. Secondly, this weight has to depend on the social characteristics of the agent speaker, i.e. tokens produced by agents that represent more innovative speakers should receive higher weights. And thirdly, the exemplar's weight should impact speech perception. That is, if a perceived token is a very atypical member of its intended (sub-)phonological category (and would therefore be rejected by the absolute memorisation criterion according to the current implementation), but is similar to stored exemplars with a high weight, it should have an increased probability of being memorised compared to an equally

[13] While in their approach a socially salient or idealised token of a word refers to "a variant or talker that is subjectively viewed as more standard compared to other variants or talkers" (Sumner et al., 2014, p. 1), it would refer to a token produced by an innovative speaker in our model.

atypical token that is acoustically similar to less heavily weighted exemplars (see Sóskuthy, 2015 for an analogous exception rule whereby exemplars can be memorised despite being phonologically ambiguous, but only if their functional load is low). These alterations to the ABM would increase the overall influence of outliers that lie in the direction of the sound change. However, it would have to be tested whether the altered version of `soundChangeR` predicts sound changes to be inevitable, which is clearly not the case: even in the face of seemingly endless phonetic variation and the presence of innovative individuals, sound changes remain rare (Weinreich et al., 1968).

4.4.2 Modelling Phonologisation

The simulation in this section failed to adequately model the change by which pre-aspiration becomes post-aspiration in clusters of /s/ plus voiceless consonant in Andalusian Spanish. In 4.3.3, this result was attributed to the lack of acoustic bias towards post-aspiration captured by the FPCA parameterisation of the data. While the parameterisation of the sounds under investigation is crucial in determining a simulation's outcome, the purpose of this section is to discuss possible shortcomings of `soundChangeR` that prevent it from modelling phonologisation not just in Andalusian Spanish, but also with regard to other cases such as vowel nasalisation (Beddor, 2009; Carignan et al., 2019; J. J. Ohala & Amador, 1981; Solé, 1995), Lausberg Italian metaphony (Greca et al., 2022; Torres-Tamarit et al., 2016), or Seoul Korean tonogenesis (Bang et al., 2018; Kang, 2014; Kang & Han, 2013).

The two key mechanisms of `soundChangeR` that can trigger phonetic and phonological changes, respectively, are the exaggeration of acoustic biases through the absolute memorisation criterion as well as the flexible and agent-specific clustering of memorised exemplars into sub-phonemic classes. Phonologisation, however, is often described in terms of cue re-weighting by listeners who perceive one cue to a contrast between speech sounds as less and another as more informative (Coetzee et al., 2018; Hagège & Haudricourt, 1978; Harmon et al., 2019; Hyman, 1976; D. Kim et al., 2017; Kirby, 2013, 2014b). Coarticulation and other biases (Garrett & Johnson, 2013) can bring about a

stable secondary cue while at the same time they can introduce variability in the primary cue, thus rendering it less reliable (Kirby, 2013). As a result, the secondary cue is eventually phonologised whereas the primary cue is weakened or vanishes entirely. Thus, it seems that implementing cues and cue weights in the ABM may be the first step towards modelling phonologisation computationally. Cues could be viewed as an abstraction over acoustic features, i.e. one or multiple acoustic features together represent one cue. Following Kirby (2013) and Toscano and McMurray (2010, 2012), cue weights can be calculated by determining how well two (sub-)phonemic classes are separated in one cue dimension compared to another. The less these classes overlap, the more informative is the cue dimension, and hence the higher is its cue weight. Since this measurement would depend on the agents' clouds of memorised exemplars in a `soundChangeR` simulation, that means that individual cue weights may differ which has proven to be the case empirically (Clayards, 2018; Idemaru et al., 2012; Kapnoula et al., 2017; Kong & Edwards, 2016; Ou et al., 2021; Schertz et al., 2015; Yu, 2021, 2022). It has also been shown that listeners can update their cue weights as a result of perceptual learning (Francis et al., 2000; Francis et al., 2008; Francis & Nusbaum, 2002; Goldstone, 1998; Harmon et al., 2019). Therefore, each agent should recompute the relative weights of each cue dimension in regular intervals, perhaps in synchrony with the recomputation of the sub-phonemic classes (i.e. right after GMM and NMF have determined new sub-phonemic classes for an agent). Importantly, this measure of informativity should impact the agents' perception given that listeners tend to categorise speech sounds by using mainly the primary, most informative cue and relying on secondary cues more heavily when the primary cue is ambiguous or unreliable (Abramson & Lisker, 1985). Even though it is unclear how exactly cues are integrated to identify the perceived sounds or words, most theoretical models of human speech perception posit that cues contribute to the categorisation of speech sounds in proportion to their weight (S. Lee & Katz, 2016; McMurray et al., 2008; McMurray & Jongman, 2011; Oden & Massaro, 1978; Toscano & McMurray, 2010, 2012). There are several possible ways of implementing perception based on cues and cue weights, the most invasive of which would be to abandon the current focus on selective memorisation

in favour of categorisation, thereby also shifting the theoretical point of view from a more exemplar-based model towards one in which errors in speech perception are a driver of sound change. More specifically, a perceived token would not be transmitted together with the word type in which it appeared, but would instead become a member of the sub-phonemic class to which it most likely belongs according to probabilistic criteria which are conditioned on the cue weights. A less radical approach would be to stick with the current criteria for selective memorisation (i.e. the intended word type is transmitted from agent speaker to agent listener), but to think of the cue dimensions as independent and integrate them only in cases of ambiguity. That is, when a perceived exemplar falls in the ambiguous space between two categories in the primary cue dimension (i.e. the one which most reliably separates two sub-phonemes from one another), it should be tested whether it passes the typicality and discriminability criteria in the secondary cue dimension. An alternative version of this idea was implemented by Kirby (2014b) in which the agents used all available cues, but disregarded a durational cue when the duration values were too low to make a difference perceptually. In general, the notion of cues is most present in studies focusing on perception and human speech recognition, but it has also been shown that an adjustment of cue weights can affect speech production (Lehet & Holt, 2017). In addition, it has been suggested that there is a correlation between cue informativity and cue enhancement (Greca et al., 2022; Kirby, 2013): if one cue is compromised (i.e. its weight decreased), another one might be enhanced proportionately by the speaker to ensure that the contrast is maintained (Cohn, 2007; Hyman, 2013). In terms of the ABM, cue enhancement can only be implemented by abandoning `soundChangeR`'s relatively narrow focus on the listener's role in sound change (Harrington et al., 2018; J. J. Ohala, 1981, 2012). In particular, the agent speaker would create a production target by sampling from a distribution over memorised exemplars of a word (as described in section 3.2.3), and then adjust that target so that it maximises the distance between the sub-phonemic classes on the cue dimension that was not compromised. This process may also occur stochastically (i.e. in any agent speaker at any time as compared to in the agent listener-turned-speaker) given that there is not

necessarily a correlation between an individual's perceptual cue weights and their reliance on the same cues in production (Lehet & Holt, 2017; Schertz et al., 2015).

Since these proposals for extending `soundChangeR` are quite pervasive and yet still underspecified from a technical perspective, it will have to be tested under which circumstances and for which kind of input data the model predicts change and stability on the phonetic, phonological, and cue levels. In any case, many questions remain to be answered with respect to phonologisation and its role in sound change, among others, why one cue should be eliminated while another is phonologised if speech sounds are redundantly cued (Lisker, 1986; Schertz & Clare, 2020), why specific secondary cues are more prone to phonologisation than others (Kirby, 2013), and how a coarticulatory effect can keep being enhanced while the coarticulatory source wanes. Computational models such as `soundChangeR` and others (Ettlinger, 2007; Kirby, 2014b; Todd et al., 2019; Wedel, 2006; Winter & Wedel, 2016) can help answer these questions, both by running simulations using models that include mechanisms proposed in the literature and by generating new testable hypotheses from these simulations for further empirical research.

5 | Conclusion

5.1 Summary

This thesis was concerned with the sound change by which pre- becomes post-aspiration in Andalusian Spanish, but also hoped to contribute more generally to our understanding of sound changes. This was attempted by presenting new approaches that take into account both the origins of sound change in the dynamics of spoken language as well as its spread through interactions between individuals.

Chapter 1 motivated the two main aims of this thesis. These were to analyse the sound change in Andalusian Spanish /sp, st, sk/ in a way that reflects its synchronic basis in gestural realignment, and to develop and demonstrate the mechanisms of an agent-based model that unifies theories on the origin and spread of sound change. For this purpose, some of the most relevant sound change theories were briefly introduced: Ohala's theory of hypo- and hyper-correction on the part of the listener, Lindblom's H&H theory that also gives an active role to the speaker, Labov's research on the social aspects of sound change, and Trudgill's mechanistic view of dialect emergence. While all of these theories, and many others, agree that sound changes arise from synchronic variability in speech production and perception, and spread through interpersonal contact, their focus clearly differs. Furthermore, none of these theories can explain why sound change happens in one language at one point in time, but not in another language or at another point in time (Weinreich et al., 1968). A holistic model of sound change that could solve the actuation problem and unify theories of origin and spread of sound changes remains to be developed. It was then proposed that computational simulations, especially agent-based models (ABMs), could provide a way of exploring the complex interplay of cognitive, social, and linguistic factors that can result in sound change. Finally, chapter 1 briefly described that /s/ before voiceless plosives is lenited in Andalusian Spanish, resulting in pre-aspirated plosives, and that this pre-aspiration is giving way to post-aspiration, thus *despide* (engl. she/he

fires) /de^{h}pide/ > /dep^{h}ide/. Previous studies have measured the duration of the aspiration phases to quantify this sound change. However, using a static measurement that relies on an artificially superimposed segmentation of the speech signal is not quite an adequate test of the underlying realignment of glottal and oral gestures which is considered the change's synchronic basis. Therefore, it was important to develop a method that could capture the dynamics of spoken language from which this sound change in Andalusian Spanish emerges.

Said method was presented in chapter 2 (which was published as Cronenberg et al., 2020). This chapter gave an overview of the historical development of pre-aspirated voiceless plosives in Andalusian Spanish before turning to a description of the articulatory bias and gestural coordination favouring post- over pre-aspiration. That is, in faster or informal speech the oral and glottal gestures needed to produce an aspirated voiceless plosive can shift from anti- to in-phase, thereby leading to an increase of post- and a decrease of pre-aspiration. This resynchronisation of gestures was modelled by means of two acoustic signals that were derived from the speech signal. The voicing probability (VP) represented the glottal opening, while the high-frequency energy signal (HF) represented the closure and friction noise. In combination, i.e. when VP was low and HF was high, these signals were taken as a measure of aspiration. HF and VP were extracted from /VsCV/ sequences (where C = /p, t, k/) produced by 48 speakers of Andalusian Spanish. Functional Principal Component Analysis (FPCA) was applied to the pairs of HF and VP signals in order to find the main dimensions of variation in their shapes. PC1 was related to the timing of the closure with respect to the glottal opening. It was furthermore shown that closure phasing impacted the extent of aspiration surrounding the closure, thus supporting the articulatory model of this sound change that claims that pre- and post-aspiration are inversely related to each other. In addition, this sound change in Andalusian Spanish was shown to be dependent on the speakers age and region: younger and West Andalusian speakers were further advanced, i.e. used more post-aspiration, than older and East Andalusian speakers. Chapter 2 suggested that two kinds of information can be extracted from the dynamic speech signal which might play import-

ant roles in both human speech processing and sound change. The first is phonological knowledge which indicates population-level characteristics of sound categories as compared to others, e.g. that /sp, st, sk/ are produced with aspiration surrounding the closure whereas /p, t, k/ are not. The second kind of information is distributional, i.e. characteristics of a speaker such as their age or regional origin correlate with the way they produce certain speech sounds. Thus, listeners who have been exposed to a wide variety of speakers are more likely to develop a perceptual trading relationship between e.g. pre- and post-aspiration which is considered a precursor to sound change.

Chapter 3 was concerned with a cognitively-inspired agent-based model of sound change that was also made publicly available as an R package called `soundChangeR`. It was argued that ABMs are very useful for the investigation of complex adaptive systems such as spoken language because they demonstrate how individual actions can lead to global patterns. The ABM `soundChangeR` is a computational implementation of the interactive-phonetic (IP) model of sound change (Harrington et al., 2018), the key components of which were introduced in detail: phonetic biases, stochastic interactions, phonetic imitation and perceptual learning, and the exemplar models of memory and phonology. Section 3.2 then explained how the IP model was implemented computationally and described the concepts, constraints, and entities that play a role in the ABM which included agents, exemplars, the production-perception feedback loop, the derivation of phonological classes, and memory management. It was concluded that sound change can emerge from the stochastic interactions between heterogeneous agents (some of which are initialised with a more innovative variant of the sound under investigation than others), given the mechanics of their production-perception feedback loop and organisation of phonological information. The core mechanisms of the model were then demonstrated in section 3.3 using two artificial datasets. The first one was used to show that the flexible phonology module of `soundChangeR` is capable of recognising systematic associations between exemplars, their location in the acoustic-phonetic space, and their word type and that, as a consequence, it identifies reasonable sub-phonemic classes. The second dataset consisted of two agents which interacted with each other, but tested the perceived tokens

either for typicality, or for discriminability, or both. These simulations resulted in phonetic shifts in different directions depending on the applied memorisation criterion, but also showed that no change emerged when none was expected. Finally, chapter 3 compared the properties of `soundChangeR` to other ABMs of sound change and discussed some opportunities to expand the model, such as the agent-specific parameterisation of data, using activation levels for memory management or for modelling lexical frequency, and introducing generational changes.

Chapter 4 combined the dynamic and computational approaches to sound change introduced in the previous two chapters and applied them to Andalusian Spanish. The aim in this chapter was to simulate the change from pre- to post-aspiration using input data that mirrors both the change (as observed in chapter 2) and the distinction between aspirated and unaspirated voiceless plosives. The inclusion of the unaspirated voiceless plosive served two purposes with regard to the simulation: it was not expected to change and the contrast between aspirated and unaspirated plosives was expected to remain stable despite the change in the aspirated ones. Thus, in section 4.2 a new dataset was composed which consisted of words containing /st/ or /t/ produced by the same 48 Andalusian Spanish speakers that provided the data for chapter 2. Once again, FPCA was applied to the voicing probability and high-frequency energy signals extracted from these data. Of the four analysed Principal Components, PC2 captured the shift from pre- to post-aspiration in /st/, which was also related to the speakers' age, whereas PC1 and PC4 captured the differences between /st/ and /t/. Although /t/ is phonologically unaspirated in Andalusian Spanish, it was produced with a small amount of post-aspiration in the analysed dataset. Nevertheless, it was shown that FPCA can derive phonological information about the variable production of voiceless plosives as suggested in chapter 2. In section 4.3, the parameterised data was submitted to the ABM `soundChangeR`. It was expected that agents representing older speakers should adopt the post-aspirated /st/ of those representing younger speakers and that /t/ should not change. Since the flexible phonology module was used, it was expected that the phonological separation between /st/ and /t/ should be maintained. While there was indeed no acoustic change

in /t/ and the phonological classes were adequately identified and maintained throughout the simulation, all agents produced the cluster with both pre- and post-aspiration by the end. This was most likely because the older agents' PC scores were not skewed towards those of the younger agents, i.e. there was no bias that could have been reinforced through the absolute memorisation criterion. In the final part of chapter 4, possible adaptations of the ABM were discussed which would enable the model to simulate sound changes in which acoustic outliers or the reweighting of perceptual cues may play a role.

In summary, this thesis presented two perspectives on the sound change by which pre- becomes post-aspiration in Andalusian Spanish /sp, st, sk/ clusters. The first perspective united the articulatory basis and cognitive implications of this change. The investigations in chapter 2 on production data from diverse speakers showed that the change towards post-aspiration came about as a result of the resynchronisation of articulatory gestures and that pre- and post-aspiration are inversely related through the relative timing of the closure. In addition, the analysis in section 4.2 supported the claim from chapter 2 that a technique like FPCA can derive phonological and distributional information from time-varying and socially variable speech data. Thus, the cognitive implications are that individuals who receive input from diverse speakers may play an important role in advancing the sound change because they are the most likely to develop a perceptual trading relationship between pre- and post-aspiration. The second perspective on the Andalusian Spanish aspiration change was computational. In section 4.3 it was shown that the computational model presented in chapter 3 was able to simulate at least parts of the sound change through stochastic interactions between heterogeneous agents who operated in an exemplar-based production-perception feedback loop and could organise and restructure their phonological classes. While the mixed results of this simulation did not generally challenge the architecture of the ABM, the simulation's shortcomings did inspire a discussion of possible expansions of the model.

5.2 Insights

The studies in this thesis have provided at least three main insights which contribute to our overall understanding of sound change. First, the analysis of time-varying acoustic signals allows for the observation of phonetic details that are relevant to sound change but might be easily missed in traditional static analyses. For instance, the study in chapter 2 showed that the amplitude of energy is an important characteristic of aspiration phases surrounding voiceless plosives in Andalusian Spanish, whereas previous studies have almost exclusively focused on the aspiration's temporal extent. Furthermore, this study was the first that was able to test the assumption that the change by which pre- becomes post-aspiration in that variety of Spanish is the result of a realignment of articulatory gestures. While previous studies attempted to frame this sound change in terms of a trade-off between the durations of pre- and post-aspiration phases, the study in chapter 2 derived acoustic signals from the speech signal that were proxies for the glottal and oral gestures involved in producing aspirated voiceless plosives. This technique was a closer approximation to the original hypothesis than static analyses of duration while it was also much less expensive and resource-intensive than collecting physiological measurements. In general, analyses that use time-varying signals can be very helpful in the investigation of sound change, because coarticulation often prevents vertical segmentations of the speech signal from being unequivocal and reliable. Moreover, the interplay between neighbouring sounds that is at the core of all sound changes can only be captured by looking beyond single segments and taking into account the dynamics of spoken language.

Second, in listener-based computational approaches to sound change, it is of utmost importance to understand how constraints on token memorisation affect simulations both in intended and unintended ways. Using simulations on an artificial dataset, it was shown in chapter 3 that the application of the absolute and relative memorisation criteria can cause very diverse simulation outcomes: while the absolute criterion can cause a shift of a broad and skewed phoneme towards a narrower one, the relative criterion leads to phoneme repulsion. However, when applied on their own, these criteria

can result in an unwanted increase or decrease of within-category variance. Therefore, it was recommended for `soundChangeR` to use both memorisation criteria so they can counteract each other's side effects. The same thorough testing approach should be taken for all relevant settings of an agent-based model. In contrast to most other computational models in this field which are usually constructed to test a specific hypothesis, `soundChangeR` has the very ambitious goal to model a wide variety of sound changes, involving either only phonetic-acoustic adjustments (such as those provoked by the memorisation criteria) or additional phonological changes. Hence, it was important to carefully test how agent-specific phonological classes are derived from remembered exemplars using the newly implemented unsupervised machine learning algorithms. In an artificially created agent, it was found that these algorithms can adequately identify sub-phonemic classes in both systematically and randomly distributed data. In addition, the study by Gubian et al. (2023) has shown that `soundChangeR` is capable of modelling phonological stability in Standard Southern British English /u/-fronting as well as the neutralisation of the phonological contrast between /ɪə, eə/ in New Zealand English.

The settings of the model are of course only one part in determining a simulation's outcome, the other part being the input data. As also stated for earlier versions of `soundChangeR` (Harrington, Gubian et al., 2019; Harrington & Schiel, 2017; Stevens et al., 2019), it is currently the only ABM of sound change that is capable of using real speech production data as input. This was very important because the sort of phonetic variation that is widely considered to be the raw material for sound changes cannot be emulated by means of artificially created data. It is for this reason that one main objective of this thesis was to develop `soundChangeR` and make it publicly available: with each new dataset that is being used in a simulation, we enhance our understanding of which mechanisms are crucial in a holistic, computational model of sound change. To my knowledge, `soundChangeR` is the first of its kind that was implemented as an R package, comes with a full documentation of the code as well as extensive demonstrations of the core mechanisms, and can be extended relatively easily if necessary. This allows all members of the research

community to test the ABM themselves and try to model the sound changes they are interested in.

Third, the study in chapter 4 combined the techniques from the previous two chapters and revealed that (i) the functional analysis of the high-frequency energy and voicing probability signals captures both the trade-off between pre- and post-aspirated /st/ and the difference between aspirated and unaspirated /t/ in Andalusian Spanish, and that (ii) the ABM `soundChangeR` failed to model the change by which pre- gives way to post-aspiration. The first part of these findings closed a desideratum from Cronenberg et al. (2020) who claimed that a data transformation such as FPCA should be able to provide functional knowledge about the phonemic contrast between /st/ and /t/. Indeed, the analysis in section 4.2 showed that there are much lower levels of aspiration in /t/ as compared to /st/ and that these differences were distributed across at least two Principal Components. Therefore, this analysis lends support for the cognitive-computational architecture in Figure 2.8. The results from the subsequent simulation on the Andalusian Spanish data, however, did not entirely align with the actual direction of change. While the separation between /st/ and /t/ was recognised and maintained by the flexible phonology module in `soundChangeR`, all agents used approximately equal amounts of pre- and post-aspiration when producing /st/ by the end of the simulation. That is, the ABM was not able to simulate the phonologisation of post-aspirated in contrast to unaspirated voiceless plosives which was expected to emerge from the interactions between younger and older agents. This result, but also the comparison to other computational models of sound change, suggested several ways in which the agent-based model could be expanded in the future.

5.3 Directions for Future Research

As mentioned in chapter 1, sound change is a complex, multi-factorial process. While a simulator must abstract from reality to achieve generalisability, there are certainly some extensions of `soundChangeR` that may be worth examining. The first is the treatment of outliers. The simulation in section 4.3 uncovered that `soundChangeR` is biased against outliers which might be a disadvantage in

modelling sound changes if exemplars of the innovative variant are especially salient, both cognitively and socially. One way to handle outliers differently would be to introduce activational weights. In exemplar theory, exemplars are weighted either with regard to the point in time of their memorisation (i.e. more recently stored exemplars have a higher weight than those stored a long time ago) or with regard to their lexical frequency (i.e. exemplars of words with higher lexical frequency receive a higher weight). Alternatively (or, possibly, in addition), it might be useful to weight exemplars according to their level of innovation, so that exemplars that lie in the direction of change and/or were produced by more progressive speakers are given a higher weight. These weights then impact both speech production and perception, so that heavily weighted exemplars have a stronger influence on the Gaussian sampling in production as exemplars with a lower weight, and can override ambiguity and atypicality issues in perception. The latter means that outliers that lie in the direction of change, but fail the memorisation criteria for being too ambiguous or too atypical, would have a chance of being memorised nevertheless.

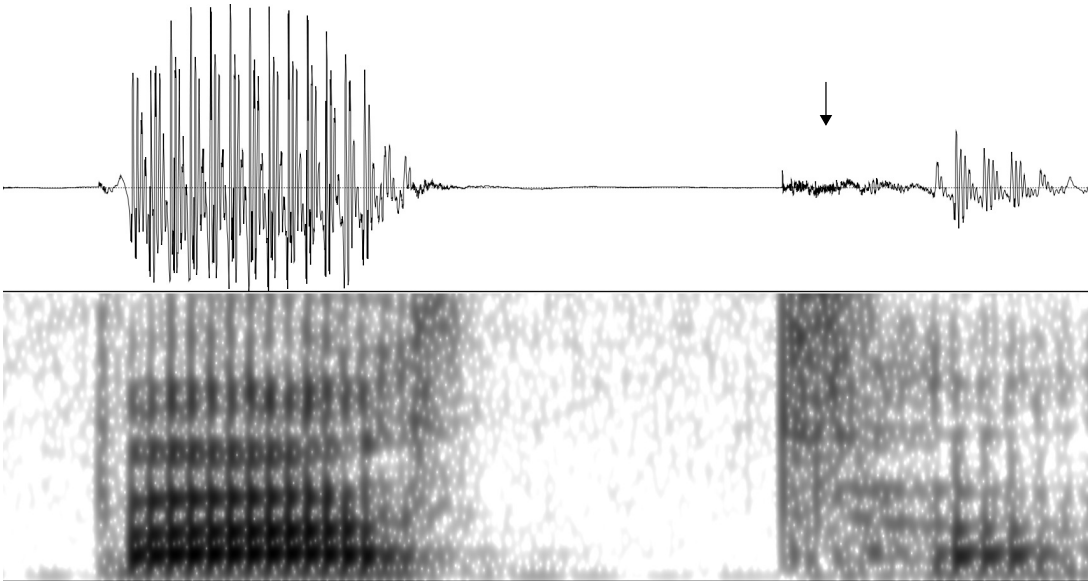
Activational levels can also fulfil different purposes in the agent-based model as discussed in chapter 3. One is memory management, i.e. avoiding an abundance of stored exemplars which would minimise the influence of newly memorised exemplars. This could be achieved by means of incrementally diminishing weights, thereby decreasing the impact of older exemplars on speech production and perception. However, this extension of the agent-based model may only counteract an issue that may not arise in reality, i.e. it is unclear how many exemplars an individual can store mentally and under which circumstances remembered traces of speech are forgotten. Somewhat more interesting would be the use of activational levels to model effects of lexical frequency on sound changes. Lexical frequency is considered influential in some sound changes, with studies showing that words change faster or slower, or are affected first or last depending on their frequency. Again, the activational weight of an exemplar, which would depend on the associated word type in this case, would influence the agents' production-perception feedback loop. Being able to model frequency effects would allow for `soundChangeR` to model a broader range of sound changes.

Another extension of `soundChangeR` that was discussed in chapter 4 given the failure to model phonologisation was the introduction of cues and cue weights. In human speech processing, cues are used to categorise speech sounds; and since such sounds can be distorted either through articulatory processes or during transmission, contrasts between speech sounds are cued redundantly, thereby facilitating accurate speech recognition. During sound changes, in particular those that involve phonologisation, the cue that most reliably signalled a phonemic contrast is rendered ambiguous while another one is enhanced and eventually takes over as the primary cue. A cue in the ABM would be a set of acoustic features which together provide information about a phonemic contrast. Thus, their weight is determined by a measure that takes into account how well two sub-phonemic classes are separated in a cue dimension and must be updated when agents have changed their sub-phonemic classes as a result of the memorisation of new exemplars. The cue weights then play a role in the agents' perception and production, although the exact implementation of their influence on these algorithms will have to be subject to testing and exploration. In perception, exemplars could be categorised or memorised according to how well they fit into one of the sub-phonemic classes in the most informative cue dimension; in this scenario, other cue dimensions are used when the exemplar is ambiguous. In production, on the other hand, the initial production target could be adjusted so that the produced exemplar is an unambiguous member of the associated sub-phonemic class in the most reliable cue dimension. All in all, cues would provide a third level besides the acoustic-phonetic and sub-phonemic level at which changes or stability can be observed. While acoustic changes are triggered by a combination of input data and the memorisation criteria, and phonological changes result from the systematicity (or lack thereof) in how exemplars are distributed in space and how they are associated to word classes, it will require thorough testing to identify the forces that impact the cue level.

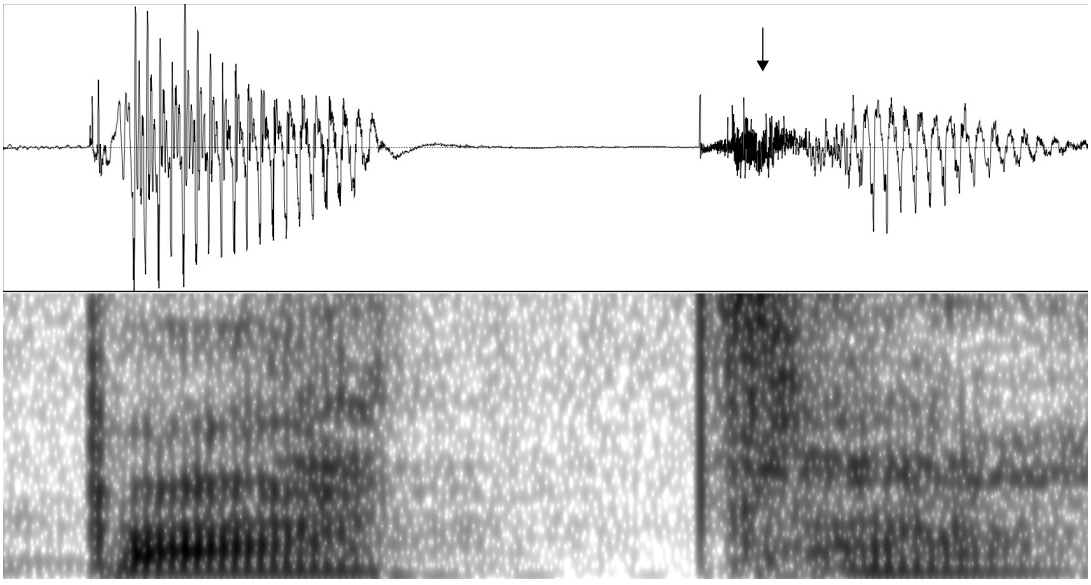
Future research should also consider the continuous nature of the aspiration change in Andalusian Spanish. According to the data presented in chapters 2 and 4, the sound change is still under way, i.e. not all speakers have adopted post-aspirated voiceless plosives and, given the trading relationship between

pre- and post-aspiration, post-aspiration has not been fully phonologised yet (Beddor, 2009). Modular feedforward models (Bermúdez-Otero & Trousdale, 2012; Fruehwald, 2017; Ramsammy, 2015) as well as the lexical phonology framework (Kiparsky, 2015) claim that sound changes proceed bottom-up, i.e. they first appear in the phonetic domain before progressing to phonology and then the lexicon. So when a sound change is still post-lexical, it applies regardless of word boundaries. In order to assess the degree of phonologisation of Andalusian Spanish post-aspiration, Egurtzegi et al. (2022) investigated whether the innovative post-aspirated variants of /p, t, k/ were also present at the word boundary, e.g. in *las tapas* /la$\underline{\text{t}}^{\text{h}}$tapah/ > [la$\underline{\text{t}}^{\text{h}}$apah] (engl. the tapas). Using the same parameterisation and analysis as in 2.2.1, it was found that younger speakers from the more progressive West Andalusian variety produced significantly more post-aspiration in /st/ across the word boundary than older speakers. However, the same was not true for /sp, sk/ across the word boundary, i.e. these clusters were realised with more pre- than post-aspiration by speakers of both age groups. In accordance with modular feedforward architecture, this finding would suggest that post-aspiration has been phonologised in /sp, sk/ ahead of /st/. However, previous research showed that /st/ was leading the sound change (Cronenberg et al., 2020; Ruch & Peters, 2016), thus posing a contradiction to the modular feedforward interpretation of the results. Further research will be needed to examine how the change from pre- to post-aspiration applies in ever narrower domains (from phonetics via phonology to lexicon) and what role the plosive's place of articulation plays.

In the data used in this thesis as well as in other recordings, it has been observed that the alveolar cluster /st/ was sometimes post-affricated instead of post-aspirated, e.g. *esto* /e$^{\text{h}}$to/ or /et$^{\text{h}}$o/ (engl. this) > /et$^{\text{s}}$o/ (Del Saz, 2019; Moya Corral, 2007; Ruch, 2010). Aside from the palatal affricate /tʃ/ as in e.g. *chica* /tʃika/ (engl. girl), there are no affricates in Spanish. It is therefore noteworthy that Andalusian Spanish has introduced /t$^{\text{s}}$/ as a variation of the post-aspirated alveolar voiceless plosive. An example of this can be seen in Figure 5.1b which shows the waveform and spectrogram of the word *pasta* (engl. pasta) produced with post-affrication as indicated by the arrow. This token can be compared to the post-aspirated token of the same word produced



(a)



(b)

Figure 5.1: Waveform and spectrogram of the word *pasta* (engl. pasta) by **(a)** an younger male West Andalusian speaker who produced the /st/ cluster with post-aspiration, and **(b)** by a younger female West Andalusian speaker who produced the cluster with post-affrication. Arrows indicate the position of the post-aspiration and post-affrication, respectively. The spectrogram range goes up to 8 kHz.

by a different Andalusian Spanish speaker in Figure 5.1a, which has a weaker release and lower centre of gravity (duration cannot be visually compared here due to different speech rates). According to a study by Ruch (2010), the dento-alveolar affricate has a similar centre of gravity as /s/ at about 6500 to 6900 Hz which is much higher than that of the glottal fricative or of post-aspiration (also see Henriksen & Harper, 2016). Del Saz (2019) showed that tokens with a centre of gravity lower than 4 kHz were clearly identifiable as /t^h/, whereas those with a centre of gravity higher than 6 kHz were identified as /t^s/. Apart from centre of gravity, acoustic measures that may distinguish post-aspiration from post-affrication include VOT, closure duration, the ratio of VOT to overall cluster duration, and zero-crossing rate (Del Saz, 2019; Ruch, 2010). Nevertheless, Ruch (2010) concluded that there was a lot of within- and between-speaker variation in how /st/ was produced and that post-aspiration and post-affrication often shared some acoustic characteristics. Therefore, it might be worth investigating in which ways /t^h/ is different from /t^s/ from both articulatory-acoustic as well as perceptual perspectives. Such investigations could also shed light on the role of affrication in the sound change that affects clusters of /s/ plus voiceless plosives in Andalusian Spanish. Relying of apparent-time data, Ruch (2010) and Vida Castro (2016) found that younger speakers used post-affrication in place of post-aspiration more so than older speakers, indicating a change in progress /t^h/ > /t^s/. In a sample of utterances from young West Andalusian speakers, Del Saz (2019) found that almost 74% of /st/ clusters were produced with post-affrication (the remaining 26% were post-aspirated). Hence, affrication might be the next step after the sound change from pre- to post-aspiration has been completed in the alveolar voiceless plosive. The current parameterisation of the change in terms of high-frequency energy and probability of voicing as explained in chapter 2, however, is not capable of capturing the difference between strong post-aspiration and post-affrication. Adding a third time-varying signal such as spectral centre of gravity or zero-crossing rate should provide the necessary distinction and could provide valuable insights into the dynamics of that sound change including both post-aspiration and post-affrication.

A | Appendices to Chapter 2

A.1 Authorship Contribution Statement

These are the authors' contributions to Cronenberg et al. (2020) according to the Contributor Role Taxonomy (CRediT). **Johanna Cronenberg**: Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Michele Gubian**: Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Jonathan Harrington**: Conceptualization, Data curation, Writing – original draft, Supervision, Project administration, Funding acquisition. **Hanna Ruch**: Resources, Data curation, Writing – review & editing.

A.2 Word List

/sp/
caspa – dandruff
espada – sword
espalda – back
España – Spain
espanto – fright
despide – he/she/it fires
espía – spy
espina – thorn
respira – he/she/it breathes
despierta – awake (fem.)
espiaba – I/he/she/it spied
espiando – spying
disputa – argument
espuma – foam
esputo – sputum
después – later

espuela – spur

/st/
vestuario – wardrobe
hasta – until
pistolín – small pistol
resto – rest
estaba – I/he/she/it was
estado – state
estanco – kiosk
pestaña – eyelash
destino – fate
estima – he/she/it respects
estío – summertime
pestiño – type of pastry
bestial – bestial
bestiando – (pseudoword)
destiempo – untimeliness
estuche – case
estufa – stove
estuve – I was
estuela – (pseudoword)
pasta – pasta

/sk/
vasca – Basque (fem.)
escama – scale
escapa – he/she/it escapes
escaso – insufficient
pescado – fish
cosquillas – tickle (noun)
esquía – he/she/it skies

esquife – skiff
esquina – corner
esquiando – skiing
escucha – he/she/it listens
escudo – shield
escupe – he/she/it spits
escuela – school
escueto – concise

A.3 Mathematical details on A_{pre} and A_{post}

In section 2.2.2 it was stated that the steepness of the curves $A_{pre}(s_1)$ and $A_{post}(s_1)$, defined in Eq. (2.2) and shown in Figure 2.5, is extreme and not due to chance, while their sum $A_{tot}(s_1)$ does not depart significantly from a flat line. Here we provide a proof of those statements.

Before proceeding with the actual proof, we apply two modifications to the definitions in Eq. (2.2) that in combination produce linear approximations of $A_{pre}(s_1)$ and $A_{post}(s_1)$. This is necessary in order to obtain properly defined slopes, and also it makes the proof more manageable. First we lift the $HF > VP$ constraint, thus including integration intervals where the area between $HF(t)$ and $VP(t)$ is negative. Then we substitute the integration boundary $t_M = \arg\min_t HF(t)$, which varies with s_1 around the middle of the time interval, with the fixed value $t = 0.5$. Looking at Figure 2.4 we can see that the impact of these modifications on the shape of $A_{pre}(s_1)$ and $A_{post}(s_1)$ will be rather modest. The inclusion of intervals where $HF < VP$ results in the inclusion of the two small intervals at the beginning and at the end of the normalised time interval where $VP(t)$ is above $HF(t)$. In those intervals, the areas delimited by the curves are small compared with the positive areas and are roughly constant at varying s_1. We can then expect that the main effect of this first approximation step is going to be a downward shift of $A_{pre}(s_1)$ and $A_{post}(s_1)$, as negative areas are going to subtract some (roughly constant) amount from the positive areas. The second approximation step is going to have a clearer, yet modest impact on the shapes of $A_{pre}(s_1)$ and $A_{post}(s_1)$, namely a mitigation

of the divergence of the two curves, especially for $s_1 > 0$. For example, in the right panel of Figure 2.4 we can see that fixing the demarcation between A_{pre} and A_{post} at $t = 0.5$ allows more area to be assigned to A_{pre} and less to A_{post}; the opposite occurs in the left panel.

Applying the above modifications to Eq. (2.2) we obtain:

$$A_{pre} \approx \int_0^{0.5} \Big(HF(t) - VP(t)\Big)\, dt \tag{A.1a}$$

$$A_{post} \approx \int_{0.5}^{1} \Big(HF(t) - VP(t)\Big)\, dt \tag{A.1b}$$

Figure A.1 compares A_{pre}, A_{post}, and A_{tot} based on Eq. (2.2), i.e. the original definition in the main text, to the approximate versions based on Eq. (A.1), where we can see that the approximate curves (solid) are indeed linear and also close enough to the original ones (dotted), especially as far as their slope is concerned. The fact that $A_{pre}(s_1)$ and $A_{post}(s_1)$ in Eq. (A.1) are linear in s_1 when $HF(t)$ and $VP(t)$ are computed only on the basis of $PC1$ follows from the linearity of FPCA and the linearity of the definite integral. To make this explicit, the following steps derive an expression for A_{pre}:

$$\begin{aligned}
A_{pre}(s_1) &\approx \int_0^{0.5} \Big(HF(t) - VP(t)\Big)\, dt \\
&= \int_0^{0.5} \Big(\mu_{HF}(t) + s_1 \cdot PC1_{HF}(t) - \mu_{VP}(t) - s_1 \cdot PC1_{VP}(t)\Big)\, dt \\
&= \underbrace{\int_0^{0.5} \Big(\mu_{HF}(t) - \mu_{VP}(t)\Big)\, dt}_{M_{pre}} + s_1 \cdot \underbrace{\int_0^{0.5} \Big(PC1_{HF}(t) - PC1_{VP}(t)\Big)\, dt}_{P_{pre}} \\
&= M_{pre} + P_{pre} \cdot s_1
\end{aligned} \tag{A.2}$$

where the first step is Eq. (A.1a), the second step is the application of Eq (2.1) using only s_1, and the rest is term rearrangement and convenient definitions of constants M_{pre} and P_{pre}. Similarly we find $A_{post}(s_1) \approx M_{post} + P_{post} \cdot s_1$ and

$A_{tot}(s_1) = A_{pre}(s_1) + A_{post}(s_1) \approx (M_{pre} + M_{post}) + (P_{pre} + P_{post}) \cdot s_1$. The four constant values are: $M_{pre} = 0.08$, $M_{post} = 0.09$, $P_{pre} = -0.21$ and $P_{post} = 0.20$.

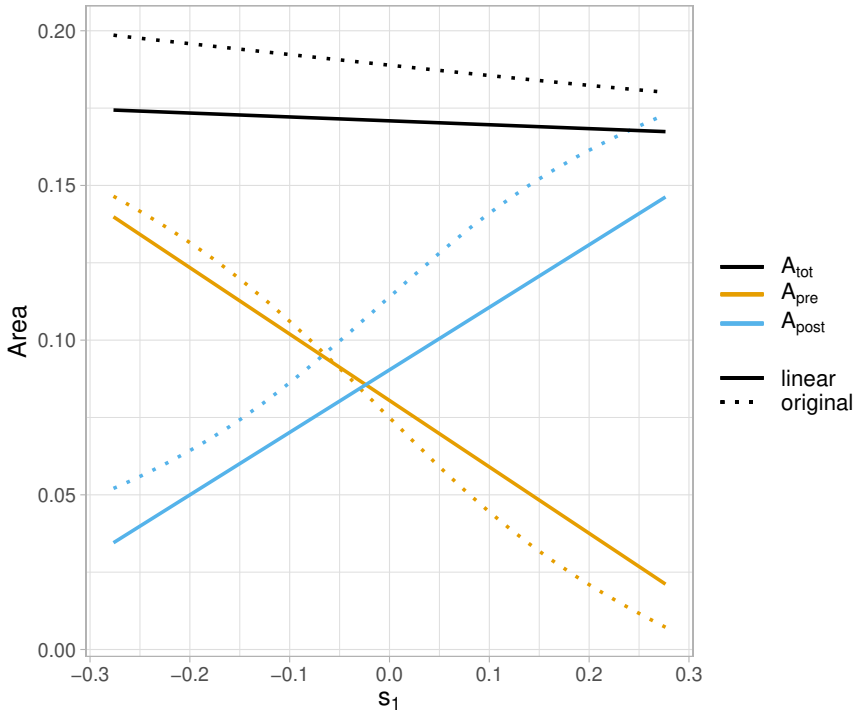


Figure A.1: A_{pre} (yellow), A_{post} (blue), and $A_{tot} = A_{pre} + A_{post}$ (black) as functions of s_1, computed by using Eq. (2.2) (dotted lines, the same as in Figure 2.5) or its approximation Eq. (A.1) (solid lines), when signals HF and VP are defined as in Eq. (2.1) using only PC1.

Having derived linear approximations to $A_{pre}(s_1)$, $A_{post}(s_1)$, and $A_{tot}(s_1)$, we focus on the corresponding slopes P_{pre}, P_{post}, and P_{tot}. We want to prove that P_{pre} and P_{post} are extreme (opposite) values, while P_{tot} is not significantly different from zero. To this end we construct a reference distribution for slopes by extending Eq. (A.1) to allow for any arbitrary partition of the time interval $(0, 1)$ in two complementary subsets on which the two integrals are computed. In this way, the partition $\{(0, 0.5), (0.5, 1)\}$ is the special case defining A_{pre} and A_{post}, while $\{(0, 1), (1, 1)\}$ defines A_{tot} (and a null area). In practice we want to show that only by setting the subdivision between A_{pre} and A_{post} in the middle of the total time interval we get a clear trade-off relationship between the two, i.e. steep and opposite slopes P_{pre} and P_{post}, while partitioning the interval

between, say, the central and remaining part or any other arbitrary partition would not provide any significant complementary relation. Operationally, we sliced the interval $(0,1)$ in $N_{int} = 20$ slots of equal size, i.e. $(0, 0.05)$, $(0.05, 0.10)$, etc., then defined the corresponding N_{int} sub-areas a_i as:

$$a_i = \int_{\frac{i-1}{N_{int}}}^{\frac{i}{N_{int}}} \Big(HF(t) - VP(t) \Big)\, dt,\ i = 1, \ldots, N_{int} \tag{A.3}$$

and then assigned a random subset of them to A'_{pre} and the remaining to A'_{post}, the generalisations of A_{pre} and A_{post} for arbitrary integration limits.

Figure A.2 shows an example of randomly partitioning the integration interval among the ones allowed by the discretisation imposed by Eq. (A.3). Each random partition will produce different $A'_{pre}(s_1)$ and $A'_{post}(s_1)$, which results in different constant terms M'_{pre}, P'_{pre}, M'_{post}, and P'_{post} defined as in Eq. (A.2). We are interested in the distribution of P'_{pre} and P'_{post}. Since those have complementary definitions, their distributions are identical, hence we will look at P'_{pre} only. There are $2^{N_{int}}$ possible partitions, that is $2^{20} = 1,048,576$ when $N_{int} = 20$, each producing a different slope P'_{pre}. We treat this large yet deterministic set of values as a population and describe it empirically as a random distribution. We estimate it by computing $10,000$ randomly chosen values of P'_{pre}.

Figure A.3 shows the Empirical Cumulative Distribution Function (ECDF) of P'_{pre}. We can immediately notice that the particular values $P_{pre} = -0.21$ and $P_{post} = 0.20$ from Eq. (A.2), which descend from (a linear approximation of) our quantification of pre- and post-aspiration, are indeed extreme (opposite) values for P'_{pre}, while their sum $P_{tot} = -0.012$ lies inside the central area of the distribution, which includes the value zero, i.e. the slope of a flat line. The distribution of P'_{pre} is quite symmetric (skewness is 0.0058), the median and mean are both around -0.005, its kurtosis is 2.78, very close to 3, which allows us to treat it as a Gaussian. The empirical quantile $q_{2.5\%}$ is -0.13, much greater than P_{pre}, and $q_{97.5\%}$ is 0.12, much smaller than P_{post}, i.e. P_{pre} and P_{post} are indeed extremes. On the other hand, the symmetric interval centred around the mean and spanning one standard deviation is $(-0.069, 0.059)$, which includes both zero and P_{tot}, hence supporting the evidence that the curve $A_{tot}(s_1)$ does

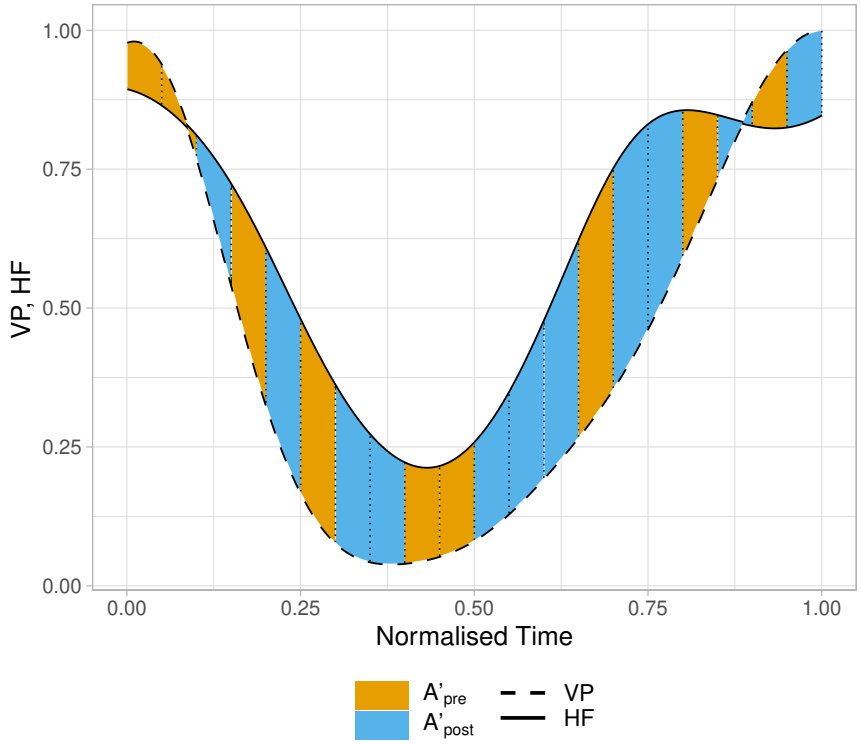


Figure A.2: Random partition of the integration interval obtained by setting $A'_{pre} = a_1 + a_2 + a_4 + a_6 + a_9 + a_{10} + a_{14} + a_{17} + a_{19}$, and $A'_{post} = A_{tot} - A'_{pre}$, where a_i's are defined in Eq. (A.3). The particular curves shown here are $HF(t) = \mu_{HF}(t)$ and $VP(t) = \mu_{VP}(t)$, i.e. $s_1 = 0$.

not depart significantly from a flat line, i.e. the total amount of aspiration is approximately constant throughout the data set we have collected.

A.4 Effects of FPCA-based signal decomposition

In section 2.2.2 it was shown for the examined data set that pre- and post-aspiration are related through the phasing of the closure when the selected acoustic signals HF and VP were derived from Eq. (2.1) using only PC1. The same relation, however, was not found when HF and VP were derived from the raw data, as shown by Figure 2.9. In this section, we will provide a more detailed explanation for this phenomenon.

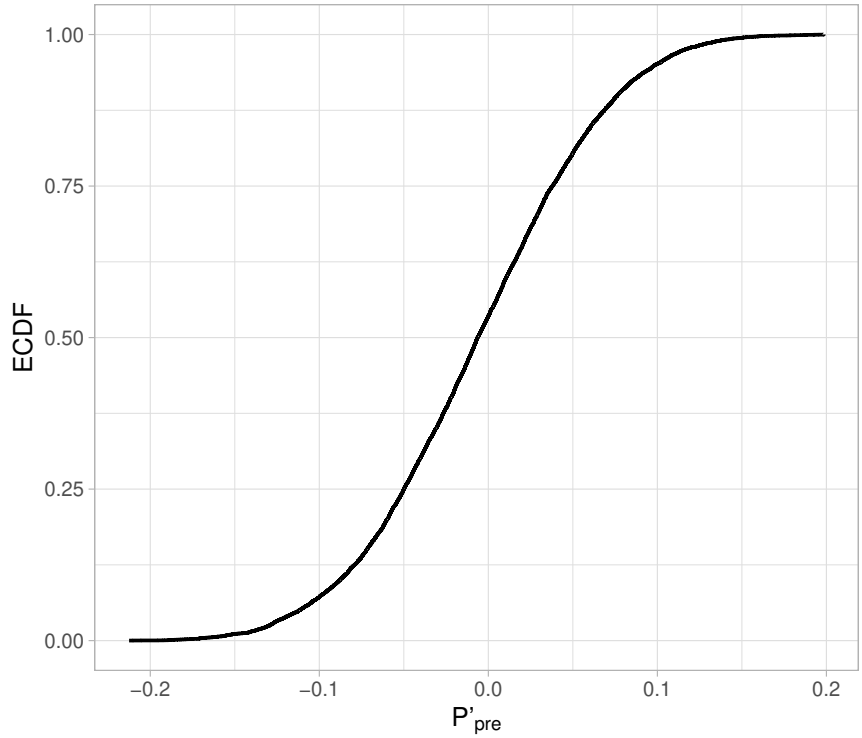


Figure A.3: Empirical Cumulative Distribution Function (ECDF) of P'_{pre}, based on 10,000 random values out of 1 million.

Acoustic signals contain a virtually infinite amount of variation, parts of which transmit attributes of the speaker to the listener. Therefore, variations that are important in a possible trade-off between pre- and post-aspiration in Andalusian Spanish can easily be masked by other sources of variation, as shown in Figure A.4. This plot was constructed in the same way as Figure 2.9, but every data point is now coloured according to its s_1 value. Recall from Figure 2.4 that positive values of s_1 were associated with earlier closures accompanied by post-aspiration, and negative values of s_1 with later closures that leave an interval for pre-aspiration. Even though the areas in Figure A.4 were calculated on the raw HF and VP signals, they bear a strong connection to the s_1 values of the data points: s_1 is positively correlated with A_{post} and negatively with A_{pre} (cf. Figure 2.5). When a token is post-aspirated according to its s_1 value (blue), it has a small A_{pre}, but can have any A_{post}, and vice versa for pre-aspirated (red) tokens. A_{post} in post-aspirated tokens can take

low or high values because of global amplitude variations in the HF signal, but importantly, A_{pre} is always small in these cases. There is, therefore, a (barely discernible) trade-off between pre- and post-aspiration in the raw data (as shown by the association of the raw data points to s_1 and consequently to the areas computed on signals reconstructed using PC1), but the trade-off is buried beneath many other kinds of variation (see e.g. Appendix A.5). This might also be the reason why other studies were unable to identify a trading relationship (e.g. Ruch, 2013; Ruch and Harrington, 2014; Ruch and Peters, 2016; Torreira, 2007).

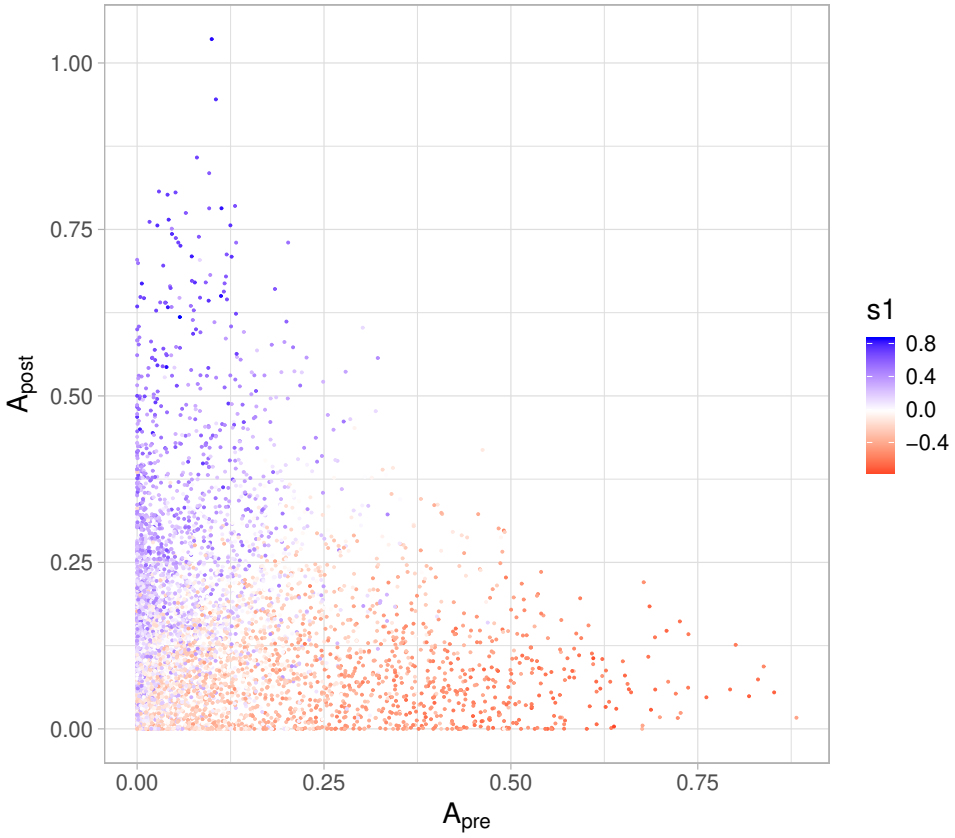


Figure A.4: Area A_{post} against A_{pre} as in Eq. (2.2) when $HF(t)$ and $VP(t)$ are the curves obtained directly from the speech signal without FPCA transformation. In comparison to Figure 2.9, the colour-coding in this plot indicated the values of PC score s_1 for each data point.

FPCA is a powerful tool for disentangling sources of variation in data sets of continuous signals so that previously unnoticeable, systematic variations can be examined while excluding other influences. This methodological advantage of FPCA is demonstrated in Figure A.5 which shows the raw data (row 2) as well as signals derived from PC1 (row 3), and both PC1 an PC2 together (row 4) for a production of *estanco* (engl. kiosk) by two different speakers. The areas A_{pre} (yellow) and A_{post} (blue) between VP and HF were computed for all panels using Eq. (2.2). The pairs of signals (and hence: the areas between them) are very similar to each other in row 3, i.e. both panels correctly show that there is more post- than pre-aspiration in the two tokens. That is because they have a very similar closure phasing (PC score s_1 is 0.34 for token 1 and 0.32 for token 2). However, when the signals are derived from both PC1 and PC2, a large difference between the two emerges: the HF signal of token 1 is shifted downwards ($s_2 = -0.12$), resulting in overall smaller areas, whereas the HF signal of token 2 is shifted upwards ($s_2 = 0.59$) which has the opposite effect on the areas. The latter signal reconstructions (row 4) are a closer approximation to the originals (row 2) than those based on PC1 only (row 3), and they still correctly indicate more post- than pre-aspiration for both tokens ($A_{post} > A_{pre}$). However, A_{post} and A_{pre} are now rendered incomparable across tokens, e.g. A_{post} in token 2 is considerably larger than A_{post} in token 1, because they are affected by a kind of variation that is irrelevant to the trade-off between pre- and post-aspiration, namely the global energy level in /sC/ clusters. From another perspective, factoring out the variation expressed by PC2 (and all other PCs except PC1) provides a way of standardising the area measurements with the consequence that the trade-off between pre- and post-aspiration emerges from the remaining amplitude differences in the PC1-derived HF signal.

A.5 Variation explained by PC2 and PC3

When applying FPCA to the input HF and VP signals, the kind of variation that is most relevant to the sound change from pre- to post-aspiration in /sC/ clusters was found to be captured by PC1. As shown in Figure 2.4, PC1 expresses a phase shift of the HF minimum relative to a voiceless interval, which

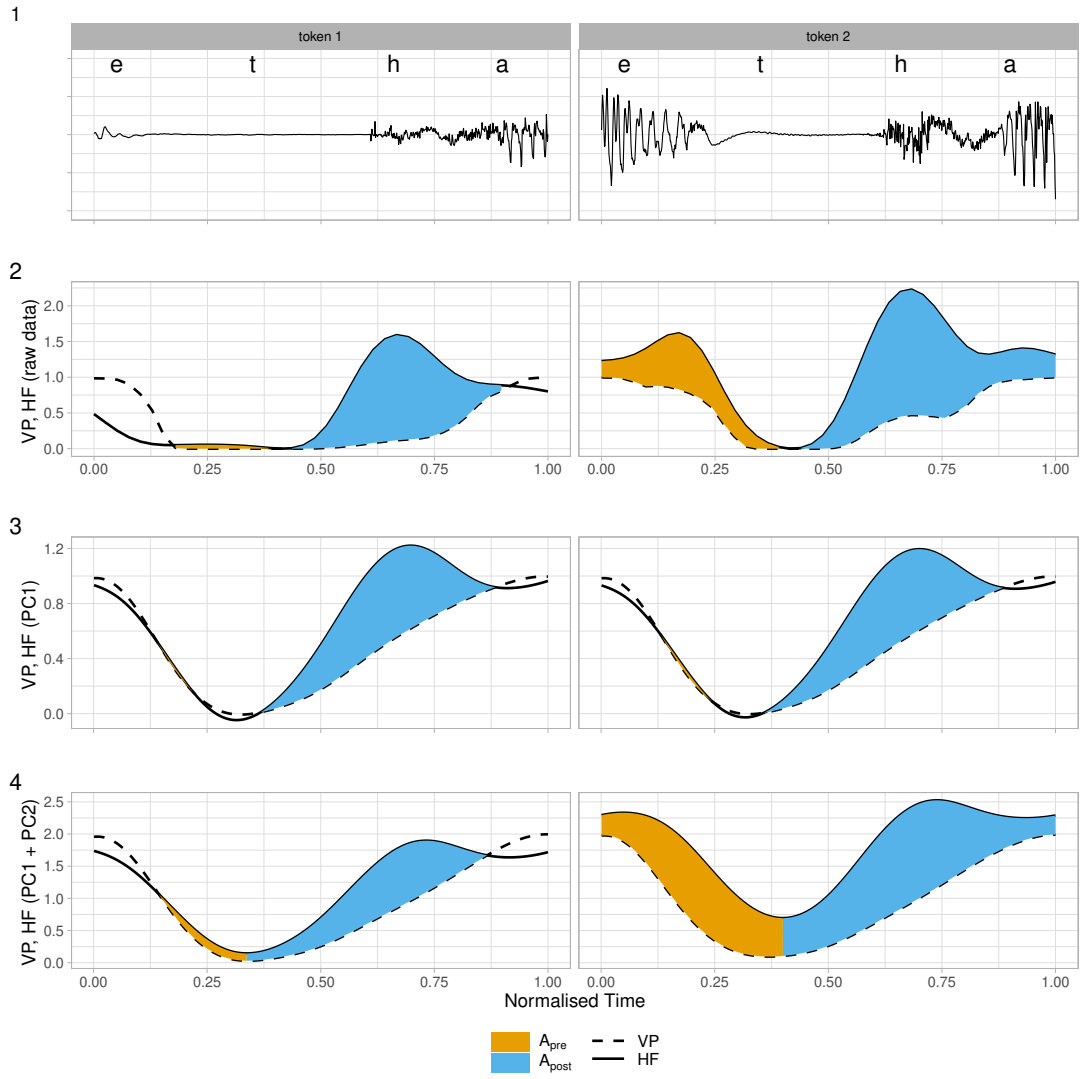


Figure A.5: Two post-aspirated tokens of the word *estanco* (engl. kiosk) produced by different young West Andalusian speakers. When reconstructing HF (solid) and VP (dashed) curves based on only PC1 (row 3), the areas A_{pre} (yellow) and A_{post} (blue) are similar for both tokens, but they are markedly different from each other when the reconstruction of the curves is based on both PC1 and PC2 (row 4).

supports the model of this sound change given in Figure 2.1 and suggests that there is a trading relationship between pre- and post-aspiration in Andalusian Spanish. However, we also computed the second and third principal component. Here we will analyse these further components and explain why we excluded them from the main part of this study.

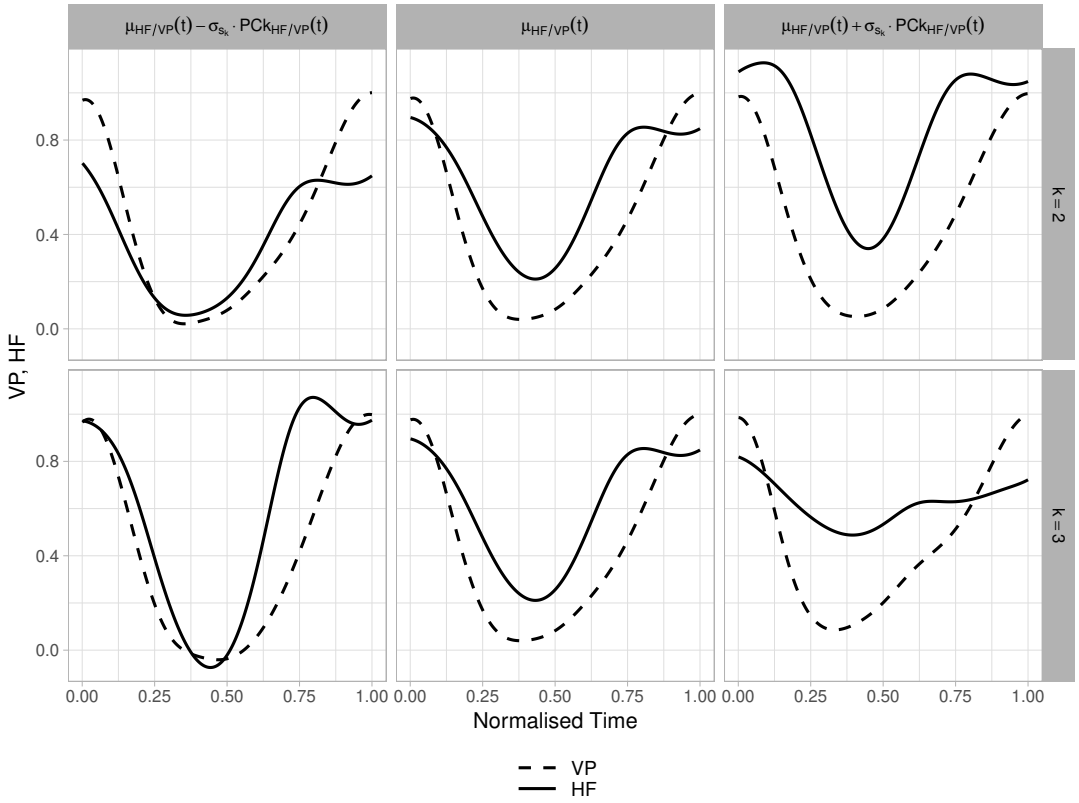


Figure A.6: Modification of the mean curves $\mu_{HF}(t)$ and $\mu_{VP}(t)$ (middle panel) by adding to (right column) or subtracting from (left column) each mean curve only one PCk curve multiplied by the standard deviation of its corresponding score (0.24 for s_2, 0.19 for s_3). The top row corresponds to PC2, the bottom row to PC3. VP curves are dashed, HF curves are solid lines.

Figure A.6 was constructed in the same way as Figure 2.4, but for the second and third PC. That is, the mean HF and VP curves are shown in the middle column; they are the same for all PCs. These mean curves were then modified by adding to (right column), or subtracting from (left column), each curve only one PCk curve multiplied by the standard deviation of its corresponding PC score (0.24 for s_2, 0.19 for s_3). The top row corresponds to PC2, the bottom row to PC3. PC2, which explained 24.3% of all variance, captures global amplitude differences predominantly in the HF signal (with almost no change to VP). This is likely caused by speaker-specific variation in energy, e.g. their distance from the microphone during the field recordings and their amplitude level

while speaking. Rather than modifying the raw speech signals, we let FPCA catch these differences. We did however try to scale the HF signals by speaker which removed most of this variation (but it did not change the main result reported in this paper), confirming that PC2 indeed captures global, speaker-specific energy levels. PC3, which explained only 14.5% of the variance in the input signals, encodes a compression and expansion of the HF curve with some slight parallel changes to VP. This kind of variation is more difficult to interpret; however, it is very clear that the PC3 curves do not contribute to an explanation of the sound change from pre- to post-aspiration in Andalusian Spanish which the present study has shown to be a phase shift of the closure in /sC/ with respect to the voiceless interval.

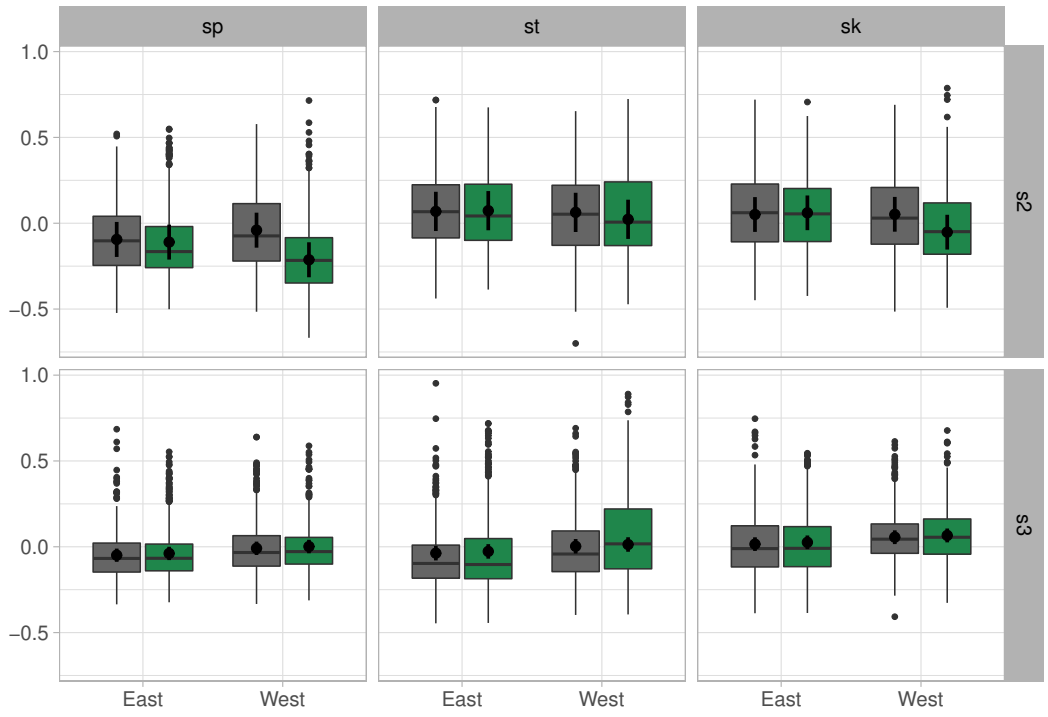


Figure A.7: Boxplots of s_2 (top row) and s_3 (bottom row) as well as their estimated marginal means (black dots within the boxes) with related confidence intervals (black vertical bands) based on the LMER models described in the text. Younger speakers are shown in green, older ones in dark grey.

We constructed the same LMER models with s_2 and s_3 as response variables as we did for s_1 in section 2.3.1. After pruning, all fixed terms with interactions were retained for s_2, whereas all interactions as well as the fixed factor age were dropped for s_3 (however, we added age back in to make the analysis homogeneous across the three PC scores). For s_2, the random intercept for word as well as the random slope for speaker were retained. All random slopes were retained for s_3. The post-hoc tests were again computed using the R package `emmeans`. This allowed us to construct Figure A.7 in the same way as Figure 2.6, but for the second and third PC scores. These boxplots show that s_2 is generally higher for alveolar and velar than for bilabial clusters. Given that PC2 captures amplitude differences in the HF signal, it is perhaps unsurprising that s_2 is lower for bilabial than for the other two cluster types as /p/ typically has a weaker burst than /t, k/. The results of the mixed model confirmed that s_2 was significantly influenced by cluster type ($F[2, 60.1] = 18.6$, $p < 0.001$). The LMER model for s_2 also shows a significant interaction between age and cluster ($F[2, 44.4] = 5.3$, $p < 0.01$) as well as a significant three-way interaction between the fixed factors ($F[2, 44.3] = 3.3$, $p < 0.05$). The former interaction can be observed in Figure A.7 where s_2 takes slightly higher values for older than for younger speakers from West Andalusia producing /sp/ or /sk/. The post-hoc tests showed that there was a significant difference between older and younger speakers from West Andalusia producing /sp/ ($t = 2.6$, $p < 0.05$). Furthermore there were significant s_2-differences between /sp/ and /st/ for older East ($t = 4.6$, $p < 0.001$), younger East ($t = 5.1$, $p < 0.001$), older West ($t = 2.9$, $p < 0.05$), and younger West Andalusian speakers ($t = 6.6$, $p < 0.001$) as well as between /sp/ and /sk/ for older East ($t = 3.9$, $p < 0.001$), younger East ($t = 4.7$, $p < 0.001$), older West ($t = 2.5$, $p < 0.05$), and younger West Andalusian speakers ($t = 4.4$, $p < 0.001$).

PC score s_3, on the other hand, was significantly influenced by region ($F[1, 50.3] = 5.8$, $p < 0.05$) and cluster type ($F[2, 65.1] = 6.2$, $p < 0.01$), as shown by the mixed model. Figure A.7 shows that s_3 was slightly higher for speakers from West than from East Andalusia for all three places of articulation, most visibly so for younger speakers producing /st/. For both age groups and all cluster types, the post-hoc tests confirmed a significant regional difference in

s_3 ($t = 2.3$, $p < 0.05$). Furthermore, there was a significant difference between /sp/ and /sk/ ($t = 3.2$, $p < 0.01$) as well as between /st/ and /sk/ ($t = 2.6$, $p < 0.05$).

Table A.1: Rounded marginal and conditional coefficients of determination for the three Linear Mixed Effects models.

PC Score	Marginal	Conditional
s_1	0.11	0.46
s_2	0.13	0.70
s_3	0.03	0.27

Table A.1 reports marginal and conditional coefficients of determination or Pseudo-R^2 scores (P. C. D. Johnson, 2014; Nakagawa & Schielzeth, 2013) for the three LMER models. These coefficients were calculated using R package `MuMIn` (version 1.43.15). The values roughly correspond to the fraction of variance explained by the fixed factors only and by the whole model, respectively. The low marginal and conditional coefficient values for PC3 indicate that this component did not contribute much to explaining the variance in the input data, as was shown previously. The random elements (Conditional - Marginal) seem to be more relevant to the s_2 model than to the s_1 model. This could indicate that the vertical shift of the high-frequency energy signal shown in the PC2 panels of Figure A.6 was conditioned more by the variation introduced by individual speakers or words, while the timing shift of the energy signal modelled by PC1 was a systematic effect governed by the fixed factors. We therefore assume PC2 to be a correcting influence on PC1 rather than to be conceptually relevant to the sound change in progress itself.

Together, the analysis of the second and third PC in this Appendix as well as Appendix A.4 show that neither of them contributes any information essential to the sound change modelled in Figure 2.1, and that both need to be factored out from the main analysis because they would otherwise obscure the relation between pre- and post-aspiration in Andalusian Spanish.

A.6 Including duration in FPCA

The analysis presented in the main text is based on the linearly time-normalised $HF(t)$ and $VP(t)$ signals, as a result of which the total duration d between time points A and B in Figure 2.2 has not been taken into account. Here we show that total duration d, though obviously varying among tokens, does not play a significant role in characterising the pre-/post-aspiration trade-off.

In order to preserve the original duration of $HF(t)$ and $VP(t)$ we adopt an extended version of multi-dimensional FPCA that incorporates time warping $r(t)$ as an extra dimension (Gubian et al., 2011). This additional curve encodes the relationship between the original and normalised time axis and decouples the information about curve shape from its duration (i.e. this is a special case of non-linear time warping). In the simple case of linear time normalisation, the time warping curve $r(t)$ is a flat horizontal line taking the value $-\log \frac{d}{\text{mean}(d)}$, i.e. (minus) the log of the normalised token duration (see Gubian et al., 2011 and Appendix A in Asano and Gubian, 2018 for an extended explanation[14]). The analysis is then carried out as a standard FPCA on the three-dimensional signals $(HF(t), VP(t), r(t))$, where the first two dimensions are the same as the ones used in the main text, thus still expressed in normalised time, while $r(t)$ separately encodes total duration in the way described above. For the analysis of the results, $r(t)$ was converted back to ordinary duration values.

Figure A.8 shows the variation of $HF(t)$ and $VP(t)$ when approximated by PC1 only. Different curve shapes are associated with different durations, from 210 ms (left panel), to 235 ms (middle panel), to 263 ms (right panel).

Despite the different type of signals, two- vs. three-dimensional, FPCA captured basically the same trends for the shape of $HF(t)$ and $VP(t)$, as can be seen by comparing Figure A.8 with Figure 2.4 in the main text, i.e. PC1 still

[14]The cited sources introduce $r(t)$ as a result of landmark registration, a procedure that was not applied here. The reader consulting those sources should consider linear time normalisation as a special case of landmark registration, where the only landmarks are placed at signal start and end. The resulting time warping function $h(t)$ is a segment whose inclination is higher (resp. lower) than 45° when total duration is higher (resp. lower) than average. The corresponding $r(t)$ is a flat line defined on the normalised time axis taking the value reported in the main text.

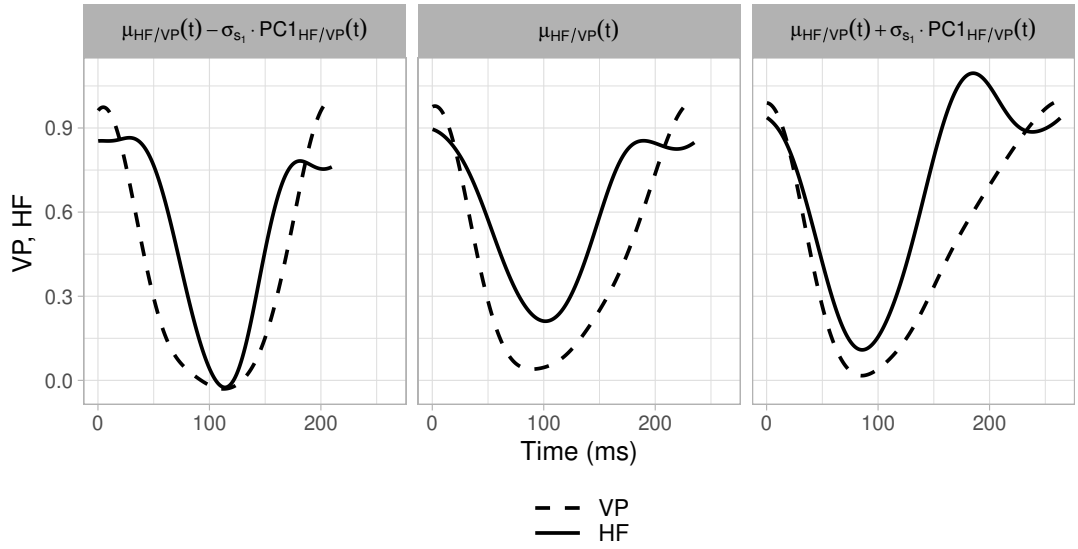


Figure A.8: Variation of $HF(t)$ (solid) and $VP(t)$ (dashed) as modulated by score s_1 (cf. Figure 2.4 in the main text). The curves were obtained applying FPCA to three-dimensional signals $(HF(t), VP(t), r(t))$, where $HF(t)$ and $VP(t)$ are in normalised time and $r(t)$ (not shown) encodes duration. The corresponding durations are, from left to right: 210 ms, 235 ms, 263 ms.

encodes a phase shift of $HF(t)$. The preservation of the trends found in the duration-agnostic analysis is not a general rule, since added information on duration can break and rearrange statistical associations (encoded by PCs). In this case, we note that longer tokens are associated with more post-aspiration (right panel), which is an expected result as post-aspiration is inherently longer than pre-aspiration.

Figure A.9 shows a corrected version of Figure 2.5, where $HF(t)$ and $VP(t)$ are approximated using PC1 from the duration-aware FPCA in which the areas were computed on unnormalised time intervals (note the different scale on the y-axis, reflecting a multiplication by duration in ms). The main difference is that A_{tot} incorporates the duration trend associated with PC1 in which a higher s_1 and hence a longer integration interval are derived from longer tokens. Despite that, the trends found in the main analysis remain, as A_{pre} and A_{post} are clearly in a trade-off, mildly distorted by the rising trend of A_{tot}. In conclusion, the duration-aware version of FPCA has enriched the analysis of

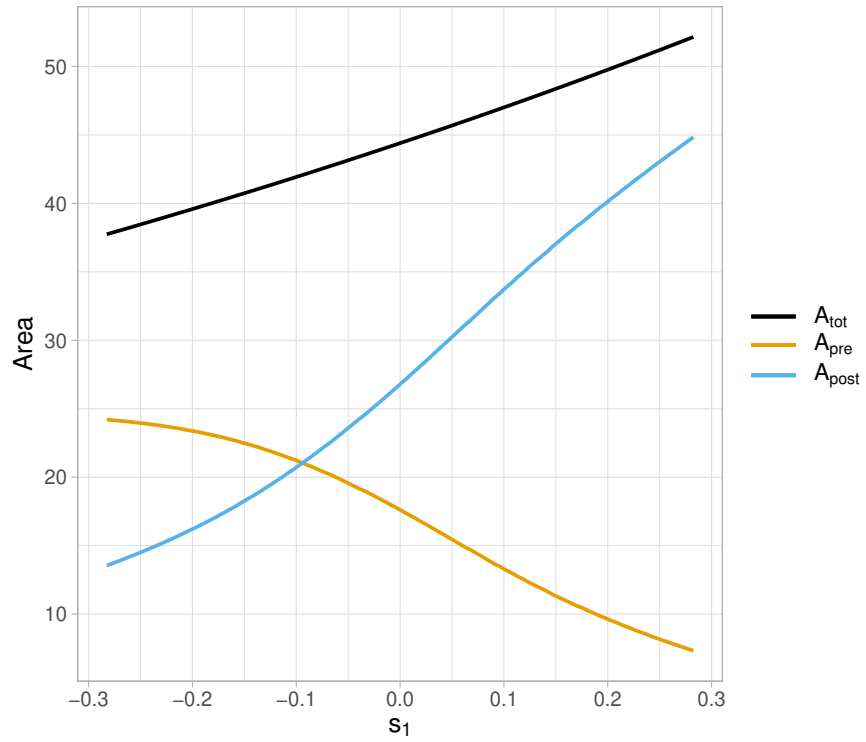


Figure A.9: A_{pre} (yellow), A_{post} (blue), and $A_{tot} = A_{pre} + A_{post}$ (black) computed as in Figure 2.5, based on $HF(t)$ and $VP(t)$ corrected for duration.

the pre-/post-aspiration trade-off reported in the main text without disrupting it in any significant way. In other words, there is no evidence that the trade-off is confounded by total duration. This is because in both FPCA analyses (with and without total duration) post-aspiration was characterised by an early closure (minimum of $HF(t)$).

A.7 Role of Areas, Time Normalisation, and FPCA

Here we illustrate how A_{pre} and A_{post} as defined in Eq. (2.2) compare to a number of alternative ways to obtain measures of pre- and post-aspiration that reliably and effectively show the underlying trade-off relation between the two. In particular, we explain the role of linear time normalisation, FPCA, as well as the use of the pre- and post-aspiration areas as opposed to using more

conventional segmental durations. For this purpose, we present a number of scatter plots showing the distribution of 5845 data points, in which each point corresponds to an /sC/ token in our data set, and the x and y axes are pre- and post-aspiration measures either in the form of areas or durations, according to different definitions and procedures. Table A.2 summarises (i) the different combinations of method of computation, either based on manual annotation or based on the $HF(t)$ and $VP(t)$ signals as defined in 2.2.1.2, (ii) whether the $HF(t)$ and $VP(t)$ signals were the raw versions or the PC1-based reconstructions, (iii) whether or not linear time normalisation was applied, and (iv) whether segmental durations or areas were used as a measure of aspiration.

Table A.2: Specifications for scatter plots in Figure A.10.

Fig.	Method	Signals	Time norm.	Measures
A.10a	manual annot.	–	no	durations
A.10b	manual annot.	–	yes	durations
A.10c	$HF(t)$, $VP(t)$	Raw	yes	durations
A.10d	$HF(t)$, $VP(t)$	PC1-based	yes	durations
A.10e	$HF(t)$, $VP(t)$	Raw	yes	areas
A.10f	$HF(t)$, $VP(t)$	PC1-based	yes	areas

Figures A.10a and A.10b are based on semi-automatically annotated durations of pre- and post-aspiration, where the annotation was taken from Ruch and Harrington (2014).[15] Figure A.10a is based on unnormalised measures, while A.10b shows the effect of dividing the durations in Figure A.10a by the total duration of the /sC/ interval (defined as in Ruch and Harrington, 2014). While pre- and post-aspiration show a mild negative correlation (Spearman's correlation are −0.13 for Figure A.10a and −0.24 for Figure A.10b), it is clear that several factors contribute to the physical duration of the acoustic manifestation of aspiration which blur the underlying trade-off that we hypothesise to be at the base of the planned articulation gesture and that we want to isolate.

Figure A.10c is obtained from the same procedure as Figure A.10b, but with the difference that the durations of pre- and post-aspiration were obtained

[15] In Figures A.10a and A.10b, d_{pre} takes negative values when the voicing of the previous vowel extended into the following closure.

from the $HF(t)$ and $VP(t)$ signals by computing the length of the integration intervals defined in Eq. (2.2). In other words, d_{pre} (resp. d_{post}) is the total duration of the interval before (resp. after) the minimum value of $HF(t)$, conditioned by $HF(t) > VP(t)$. There is an improvement in the visible correlation (Spearman's correlation is -0.31), but still the trade-off is not clearly delineated from other factors that contribute to duration. The situation is even worse when computing the areas A_{pre} and A_{post} instead of durations, as shown in Figure A.10e (same as Figure 2.9). By contrast, a few lines rather than a cloud of points are obtained when the same parameters were derived from the PC1-based reconstructed signals (Figures A.10d and A.10f). Those lines are still scatter plots, i.e. formed by individual points, but this time the location of the points is constrained by a single degree of freedom, i.e. PC score s_1. In Figure A.10d a large portion of the scatter plot exhibits an obvious trade-off (the segment with roughly $-45°$ inclination), while other parts are affected by what we argue are artefacts. These are the consequence of the fact that $d_{pre} + d_{post} \leq 1$ (because the total duration of the signals is 1 in normalised time), but at the same time d_{pre} (resp. d_{post}) cannot be larger than t_M (resp. $1 - t_M$), where t_M is the location of the minimum of $HF(t)$, which is usually around the temporal midpoint $t = 0.5$ and it rarely occurs near $t = 0$ or $t = 1$. As a consequence, when either d_{pre} or d_{post} decreases below approx. 0.3, the other stops increasing, i.e. the trade-off between d_{pre} and d_{post} is interrupted. This explains e.g. the roughly vertical line at the bottom right corner in Figure A.10d, where d_{post} keeps decreasing from 0.3 to 0.2 while d_{pre} stops increasing (and even decreases slightly), as according to its definition its value cannot exceed t_M, which is not likely to be far from $t = 0.5$. These artefacts are not present when areas were used instead of durations, as Figure A.10f illustrates, where a clear trade-off relation is preserved even when either A_{pre} or A_{post} are small.

To summarise, with Figure A.10 we have shown that for the purpose of isolating the underlying pre-/post-aspiration trade-off (i) linear time normalisation alone does not bring any particular benefit, (ii) a clear trade-off emerges only by applying the FPCA-based signal decomposition on $HF(t)$ and $VP(t)$, and when doing so, (iii) computing areas instead of segmental durations

preserves a clear trade-off trend also at the extremes of pre-/post-aspiration ranges.

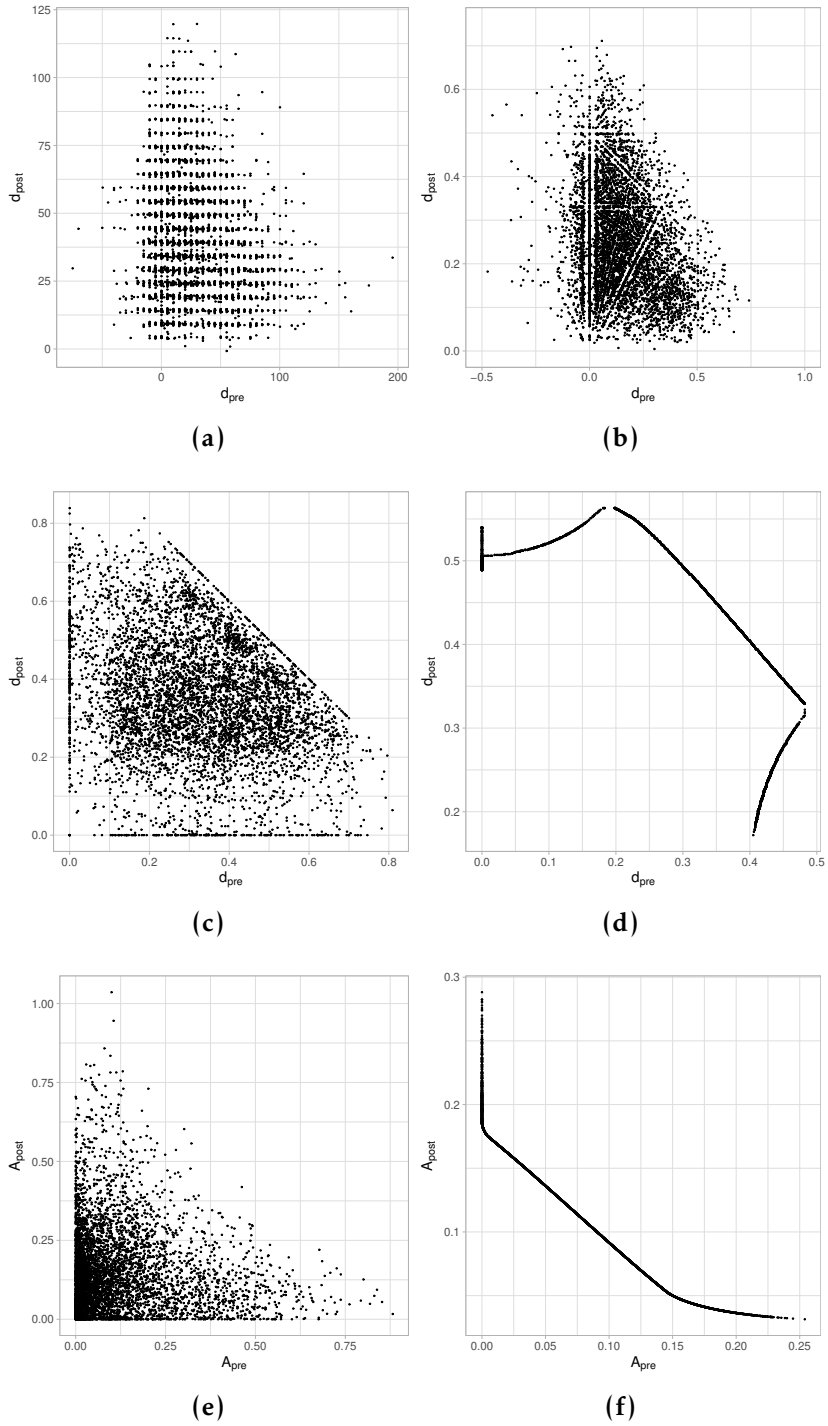


Figure A.10: Scatter plots showing the distribution of pre- (x-axis) and post-aspiration (y-axis) measures obtained using different methods. See Table A.2 and text for details.

B | Appendices to Chapter 3

B.1 Authorship Contribution Statement

These are the authors' contributions to Gubian et al. (2023) according to the Contributor Role Taxonomy (CRediT). **Michele Gubian**: Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Johanna Cronenberg**: Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jonathan Harrington**: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. Please note that Appendices B.4 to B.7 are largely the work of Michele Gubian and have been taken from Gubian et al. (2023) with some rearrangements and adaptations to make them align with the contents of chapter 3.

B.2 Installation of soundChangeR

The R package `soundChangeR` can be installed via GitHub (i.e. it is not available on CRAN). The link to the repository of `soundChangeR` is https://github.com/IPS-LMU/soundChangeR and the most recent version of the package will always be on the branch called `main`. At the time of writing, the current version of `soundChangeR` is 1.0.0. As usual before installing a new package in R, please make sure all of your packages are up-to-date and you are on an R version equal to or higher than 4.1.0. If necessary, install `devtools` first, before installing `soundChangeR`:

```
install.packages("devtools")
devtools::install_github("https://github.com/IPS-LMU/
    soundChangeR", build_vignettes = T)
```

There is a vignette available for this package which gives an overview of the mechanics of the model and explains each parameter that can be set in a simulation (see Appendix B.3).

B.3 Vignette to soundChangeR

Simulating Sound Changes with soundChangeR

Johanna Cronenberg

2022-09-15

- 1 Interactions
  - 1.1 Production
  - 1.2 Perception
  - 1.3 Phonology
  - 1.4 Memory Management
- 2 Parameters of the Model
  - 2.1 Input data
  - 2.2 Setup
  - 2.3 Production
  - 2.4 Perception
  - 2.5 Phonology
  - 2.6 Interactions
  - 2.7 Runs
  - 2.8 Other options
- 3 Analysing Results and Managing Simulations
  - 3.1 Demo
  - 3.2 Data Structures
  - 3.3 Analysis of Results
  - 3.4 Managing Simulations
- 4 Recommended Literature

`soundChangeR` is an agent-based computational implementation of the interactive-phonetic (IP) model of sound change (Harrington et al., 2018). The central entities in this model are agents and exemplars. Agents are artificial representations of real speakers, i.e. the agents are endowed with a memory filled with acoustic data from actual speakers. The agents also follow a set of rules in order to produce, perceive, and forget exemplars. Exemplars consist of acoustic parameters that capture essential characteristics of the speech sounds under investigation. Every exemplar is also associated to a fixed lexical class, i.e. the word in which the exemplar was produced. The phonemic level of this model links the acoustic exemplars and the word classes. Sound change may or may not emerge from the interactions between the agents as well as the mechanisms of their production-perception feedback loop.

We will explain the ABM's mechanisms by means of the exemplary dataset which is provided with the package:

```
library(soundChangeR)
u_fronting
#> # A tibble: 2,390 × 8
#>    sl_rowIdx speaker age     word  label  DCT0  DCT1   DCT2
#>        <int> <chr>   <chr>   <chr> <chr> <dbl> <dbl>  <dbl>
#>  1         1 albr    younger seep  i:    1453. -20.8 -14.4
#>  2         2 albr    younger seep  i:    1438. -35.2 -24.8
#>  3         3 albr    younger seep  i:    1419. -34.6 -21.0
#>  4         4 albr    younger seep  i:    1438. -27.3 -13.6
#>  5         5 albr    younger seep  i:    1443. -27.5 -11.9
#>  6         6 albr    younger seep  i:    1423. -29.2 -14.5
#>  7         7 albr    younger seep  i:    1432. -14.3 -13.4
#>  8         8 albr    younger seep  i:    1449. -32.7 -14.0
#>  9         9 albr    younger seep  i:    1450. -31.7 -19.2
#> 10        10 albr    younger seep  i:    1438. -34.6  -8.71
#> # … with 2,380 more rows
```

The dataset `u_fronting` contains data from 22 speakers of Standard Southern British English (SSBE) producing the three vowels /i, u, ju/ in a total of 11 words. Each word was repeated usually 10 times per speaker. The vowels are parameterised by means of the first three DCT coefficients of the second formant trajectory. A help page for this dataset is available:

```
?u_fronting
```

More detailed information on the data is provided by Harrington & Schiel, 2017.

1 Interactions

During a simulation with `soundChangeR`, the agents exchange tokens for a given amount of interactions. An interaction always consists of an agent speaker who produces a new token of a word and an agent listener who decides whether or not to memorise the token. Agents can interact with each other either freely or they do so only within or only across predefined groups (such as older vs. younger, region A vs. region B, etc.).

In `u_fronting`, the agent groups are based on the speakers' age:

```
unique(u_fronting$age)
#> [1] "younger" "older"
```

This is because empirical studies have found that younger speakers of SSBE are further advanced in the change of /u/ to the front of the vowel space than older speakers. So, whereas younger speakers have a /u, ju/ close to /i/, older speakers still produce a retracted /u, ju/ most of the time. However, when /u/ appears in adjacency to coronal consonants, even older speakers are more likely to produce a fronted /u/ due to the high F2-locus of coronal consontants. This can be shown by calculating the mean DCT0 coefficient by vowel and age group in `u_fronting`:

```
#> # A tibble: 6 × 3
#>   label age     mean_DCT0
#>   <chr> <chr>       <dbl>
#> 1 i:    older       1416.
```

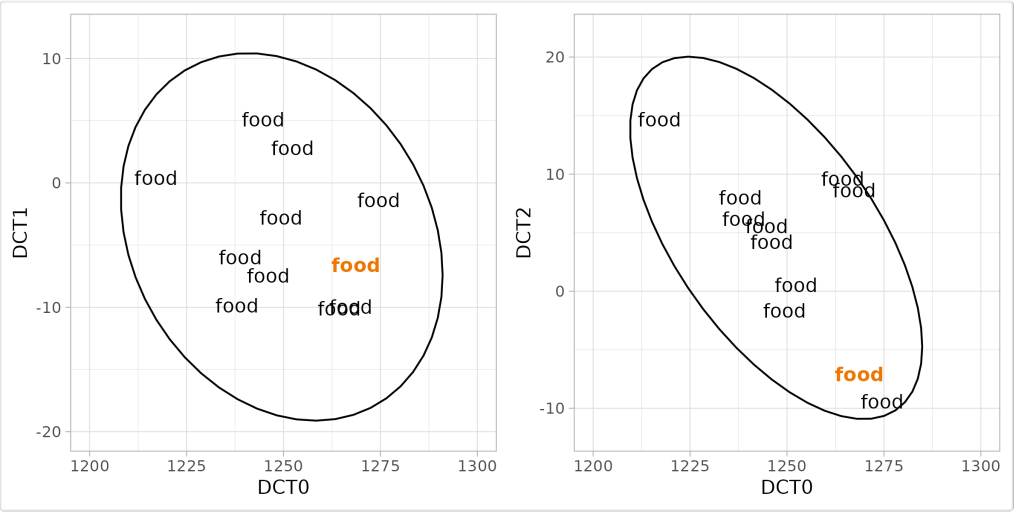
```
#> 2 i:    younger     1469.
#> 3 ju:   older       1099.
#> 4 ju:   younger     1377.
#> 5 u:    older        926.
#> 6 u:    younger     1299.
```

DCT0 is linearly related to the mean of the F2 trajectory, i.e. the higher DCT0, the higher F2. While there is hardly any difference in the mean DCT0 of /i/ between older and younger speakers, older speakers have much lower DCT0 values for /u, ju/ than younger ones.

1.1 Production

The agent speaker randomly chooses a word class, then builds a Gaussian model over all memorised exemplars associated with that word, and samples a new token from it. This process is word-based in order to ensure that possible coarticulatory effects can be carried over into the new token.

Say, `albr` is the agent speaker and has chosen to produce a token of the word *food*. Then `albr` gathers all exemplars of *food* (as shown by the `food` labels in the plots below) and estimates a Gaussian model over them (as exemplified by the ellipses) in the three-dimensional DCT space (here broken down into two 2D plots for reasons of legibility). A new token of *food*, shown in orange, is then sampled from the Gaussian distribution:



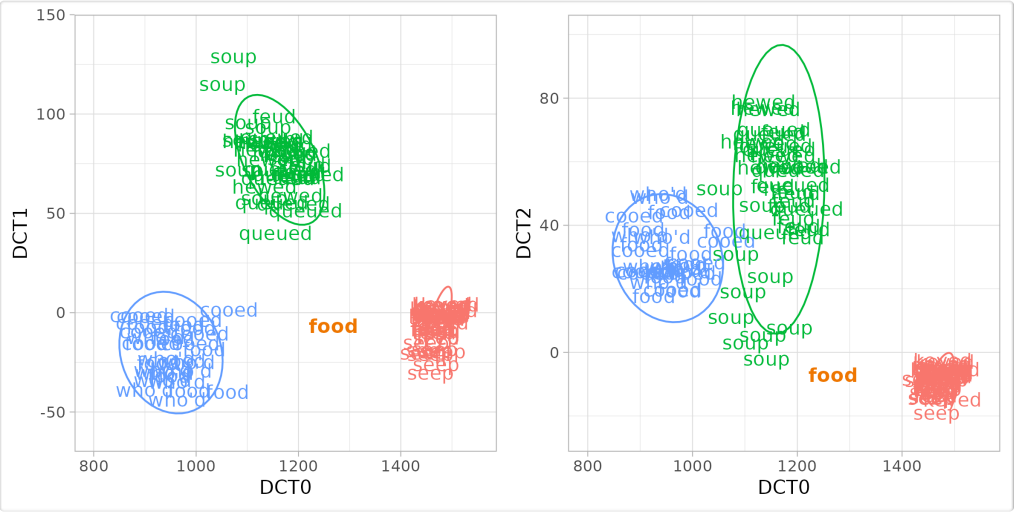
The new token consists of the values for the three DCT coefficients as well as the label of the word class, *food*.

1.2 Perception

The agent listener receives the token together with its associated word class. This means that it is assumed that word recognition works perfectly and that misunderstandings are neither a catalyst nor an obstacle for sound change. Instead, the agent listener has to decide whether or not to *memorise* the token. This decision is strongly linked to the phonemic level, i.e. the agent tests whether the perceived token is close enough to the intended phonemic class and/or probabilistically closer to the intended than to all other phonemic classes. The phonemic classes can either be pre-determined by the user and remain fixed or be regularly updated by

each agent by means of unsupervised learning algorithms.

Continuing with the example above, say, `elwi` is the agent listener who has to decide whether or not to memorise the perceived token of *food*. `elwi`'s acoustic space with the phonemic categories colour-coded is shown in the following plot. The perceived token of *food* is again in orange.

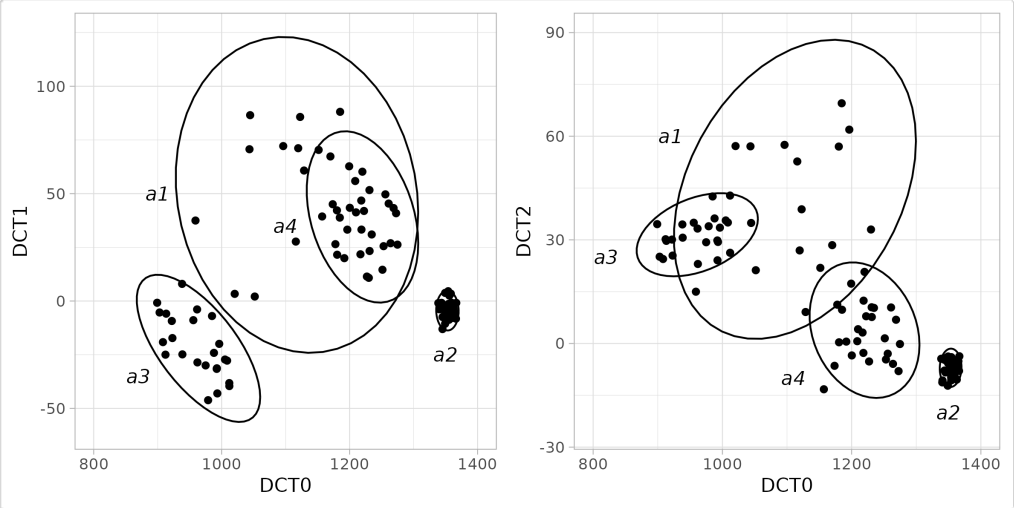


For `elwi`, *food* is associated with the blue phonemic class. `elwi` can either accept the token without any constraint, or use different probabilistic decisions (also see memoryIntakeStrategy). One of them tests whether the token of *food* is close enough to its intended phonemic class without taking any other phonemic classes into account. If, according to the Mahalanobis distance, the token is too far from the blue phonemic class, `elwi` rejects the token. The other default decision metric is the maximum posterior probability criterion which takes all phonemic classes into account. The new token of *food* must be probabilistically closer to the blue than to all other phonemic classes. If this is the case, `elwi` accepts and memorises the token. In the given example, both probabilistic decisions would lead to a rejection of the token.

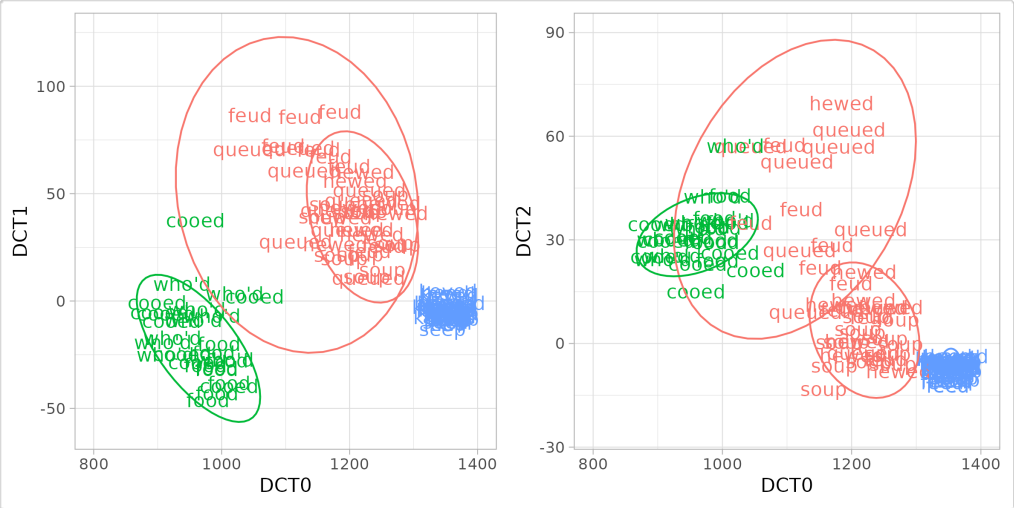
1.3 Phonology

There are two ways of linking the exemplars and word classes through a phonemic level. Either the phonemic classes are fixed and immutable throughout the simulation (like in the demo simulation) or they are agent-specific, regularly updated, and computed using unsupervised machine learning mechanisms (see useFlexiblePhonology). The latter option consists of a two-step process: First, Gaussian Mixture Models (GMM) are used to create acoustic clusters of exemplars. This step relies exclusively on information about the location of exemplars in the acoustic space (and no information about word classes etc.).

The plot below shows the components of the GMM for the speaker `phfo` from a simulation with the `u_fronting` dataset after 100,000 interactions. Every black dot is an exemplar in `phfo`'s memory at that point in the simulation. At this first stage of deriving phonemic knowledge, the exemplars' association to word classes is irrelevant, thus only the acoustic information is used to form clusters. The GMM has determined that there are four acoustic clusters (*a1* to *a4*) as shown by the labelled ellipses.



In the second step, non-negative matrix factorisation (NMF) is used to identify sets of acoustic clusters that contain exemplars of the same distinct word classes. These sets are called sub-phonemes. This step neglects any information on the location of exemplars and acoustic clusters in the acoustic space, and instead uses information on the association between exemplars and word classes.



Therefore, every exemplar from the previous plot is now represented by the associated word label. In total, NMF has determined that there are three sub-phonemes as shown by the different colours. Acoustic clusters *a1* and *a4* both contain mostly exemplars of the same four word classes: *queued*, *feud*, *hewed*, and *soup*. That is why *a1* and *a4* are grouped together into the red sub-phoneme. Most exemplars of the words *cooed*, *food*, and *who'd* are contained in acoustic cluster *a3* and no other cluster, hence *a3* becomes the green sub-phoneme. However, two exemplars of *cooed* and one exemplar of *who'd* were originally part of acoustic cluster *a1*. These exemplars are so-called impurities in the red sub-phoneme (with which cluster *a1* is associated). Sub-phonemes can contain impurities as long as the overall purity of the sub-phoneme surpasses a given threshold. Purity is computed as the fraction of exemplars in a sub-phoneme belonging to a designated set of words. So if the red sub-phoneme consists of a total of 27 exemplars, 24 of which are associated with the word classes *queued*, *feud*, *hewed*,

and *soup*, then the purity is

```
24/27
#> [1] 0.8888889
```

which is higher than the default purity threshold of 0.75. Finally, acoustic cluster *a2* exclusively contains exemplars of the words *feed*, *heed*, *keyed*, and *seep* and is hence identified as the blue sub-phoneme.

The sub-phonemic classes are recomputed at regular intervals. Importantly, previous results from GMM and NMF are disregarded and play no role in the recomputation of the sub-phonemes. If this flexible phonology module is used to derive phonemic knowledge, the memorisation criteria are computed with respect to these sub-phonemes.

1.4 Memory Management

There are two scenarios to be avoided in terms of memory size: When there are too few exemplars per word and agent, Gaussian distributions (as computed in production) are unstable or cannot be computed; when there are too many exemplars in the agents' memories, the influence of new exemplars is minimised and change is effectively inhibited. To solve the first issue, the user can choose to apply SMOTE, both at initialisation and during production. SMOTE is a standard resampling algorithm that has no harmful effects on the acoustic distributions.

The second issue (i.e. too many exemplars) can be avoided by having the agent listeners forget exemplars, i.e. remove an exemplar from memory that belongs to the same word class as the memorised token. The only restriction on forgetting is that word classes cannot be diminished, so if the deletion of an exemplar would leave the word class with less exemplars than it was initialised with, forgetting is blocked.

Let's assume that `elwi` has decided to memorise the perceived token of *food* in the example above. Depending on the model's settings, a random exemplar of *food* can be removed unless the result is that there are less than 10 exemplars of *food* left afterwards. This is because `elwi` was initialised with 10 exemplars of *food*.

2 Parameters of the Model

The function which starts a simulation is called `run_simulation()`. The following is a comprehensive list of all parameters of the model that are arguments of `run_simulation()`. Short explanations of the arguments are also available on the function's help page:

```
?run_simulation
```

2.1 Input data

```
inputDataFile = NULL
speaker = NULL
group = NULL
word = NULL
phoneme = NULL
features = NULL
```

The first five arguments all take strings or a vector of strings. `inputDataFile` is the relative or absolute path to a data file, which can be a `.csv` or simple `.txt` file. When running the simulation, it will be loaded as a `data.table`. The argument `speaker` indicates the column in `inputDataFile` which contains the speaker codes. The argument `group` can be used if there are two or more agent groups, based on e.g. age or regional origin. If there are no groups, this argument can remain `NULL`, otherwise it is the name of the respective column in `inputDataFile`. The argument `word` indicates the column in `inputDataFile` which contains the lexical labels. If you are not using the flexible phonology algorithm (i.e. if `useFlexiblePhonology = FALSE`), the argument `phoneme` has to be specified and must point to the column in `inputDataFile` that contains the canonical phonemic labels. Otherwise, `phoneme` can remain `NULL`. For argument `features`, please indicate the name(s) of the column(s) that contain the acoustic parameter(s) (formant values, DCT coefficients, PC scores, etc.) of the sounds under investigation. The indicated column(s) must contain numbers. There can be more columns in `inputDataFile` than needed for the simulation (they will be ignored).

An example for these arguments can be given using the data frame `u_fronting` again. This dataset is also used in the demo of this model for which the arguments above are set as follows:

```
speaker = "speaker"
group = "age"
word = "word"
phoneme = "label"
features = c("DCT0", "DCT1", "DCT2")
```

The argument `inputDataFile` in this case points to the `.csv` file of `u_fronting` which is located at:

```
system.file("extdata", "u_fronting.csv", package = "soundChangeR")
```

2.2 Setup

```
subsetSpeakers = NULL
subsetPhonemes = NULL
createBootstrappedPopulation = FALSE
bootstrapPopulationSize = 50
expandMemory = FALSE
expandMemoryFactor = 2
removeOriginalExemplars = FALSE
```

If only a subset of speakers or a subset of canonical phonemes should be used in the

simulation, the arguments `subsetSpeakers` and `subsetPhonemes` can be set with vectors of strings containing the speaker codes or phonemes, respectively, that are to be part of the simulation.

Usually, every human speaker is represented by one agent which is achieved by initialising the agent with the acoustic data of the speaker (`createBootstrappedPopulation = FALSE`). However, there is the possibility to break this paradigm by applying bootstrapping. If so, the argument `bootstrapPopulationSize` must specify how many agents (per agent group) the new population should be comprised of (a full positive number). For example, the `u_fronting` dataset consists of data from 11 older and 11 younger speakers. If the bootstrapped population should consist of a total of 50 agents (irregardless of their age groups), the arguments need to be:

```
createBootstrappedPopulation = TRUE
bootstrapPopulationSize = 50
```

If, instead, the bootstrapped population should consist of 25 older and 25 younger agents, the arguments are:

```
createBootstrappedPopulation = TRUE
bootstrapPopulationSize = c("older" = 25, "younger" = 25)
```

It is highly recommended to do multiple runs of simulations when the population is created by means of bootstrapping.

Since we often deal with sparse data in the phonetic sciences, it makes sense to augment the amount of data (i.e. tokens per word per speaker) *before* the first interaction takes place by setting `expandMemory = TRUE`. In this case, `expandMemoryFactor` needs to be set to a full positive number to indicate the factor by which to multiply the number of tokens per word and speaker. The memory expansion uses the production technique indicated by the parameters in the next section, i.e. the agent is essentially talking to itself to create more tokens. Be aware that a large expansion factor, e.g. 10, will slow any change down, so many more interactions are needed. The argument `removeOriginalExemplars` only takes effect if `expandMemory = TRUE` and results in the deletion of all original exemplars from the agents' memories before the interactions begin.

2.3 Production

```
useSMOTE = TRUE
fallBackOnPhoneme = TRUE
minTokens = 10
SMOTENN = 5
```

The agent speaker randomly chooses a word class from which to produce a new token. If the number of exemplars associated with the chosen word class is less than `minTokens`, SMOTE can be applied to temporarily create more tokens of the word by setting `useSMOTE = TRUE`. This makes the estimation of the (often multi-dimensional) Gaussian distribution from which a new token is sampled more stable. In this case, `SMOTENN` specifies the number of nearest neighbours to be considered when performing the random linear interpolation for SMOTE.

Before applying SMOTE, one can decide to additionally use exemplars of the same phonemic

class with which the chosen word is associated. So if `fallBackOnPhoneme = TRUE`, the exemplars associated with the chosen word as well as exemplars of the same phonemic class are used to estimate the Gaussian distribution. Only if these exemplars are still less than `minTokens`, SMOTE is applied. If `fallBackOnPhoneme = FALSE`, SMOTE is used immediately. If `useSMOTE = FALSE`, the argument `fallBackOnPhoneme` takes no effect.

2.4 Perception

```
memoryIntakeStrategy = c("mahalanobisDistance", "maxPosteriorProb")
mahalanobisProbThreshold = .95
posteriorProbThreshold = 1/3
perceptionOOVNN = 5
forgettingRate = 1
```

The agent listener has several options for deciding whether or not to memorise the perceived token. The argument `memoryIntakeStrategy` is therefore one of the most critical ones in this model and can take either a string or vector of strings as values (i.e. combining strategies is possible).

- `"mahalanobisDistance"`: The distance between the token and the corresponding phonemic class in the agent listener's memory has to be smaller than the `mahalanobisProbThreshold` which takes a value between 0 and 1. This approach does not take into account any of the other phonemic categories.
- `"maxPosteriorProb"`: Maximum posterior probability decision, i.e. the token is only memorised if its probability of belonging to the listener's corresponding phonemic category is higher than that of belonging to any of the other categories.
- `"posteriorProbThr"`: The produced token is memorised if its posterior probability of belonging to the phonemic category is higher than the threshold indicated by the argument `posteriorProbThreshold`.
- `"acceptAll"`: All perceived tokens are also memorised, i.e. there is no constraint on memorisation.

The two possible combinations of values for `memoryIntakeStrategy` are:

```
memoryIntakeStrategy = c("mahalanobisDistance", "maxPosteriorProb")
memoryIntakeStrategy = c("mahalanobisDistance", "posteriorProbThr")
```

If the perceived token is associated with a word that is unknown to an agent listener, a word label will be assigned to the token based on a majority vote among `perceptionOOVNN` nearest neighbours. This argument therefore has to be an uneven full positive number.

If the perceived token has been memorised, the agent listener can forget an exemplar of the same word class if a value sampled from a uniform distribution is below `forgettingRate`. So the agent listener will always remove an exemplar if `forgettingRate = 1`, never remove one if `forgettingRate = 0`, and remove a token some of the time if `forgettingRate` is set to a value between 0 and 1.

2.5 Phonology

```
useFlexiblePhonology = FALSE
computeGMMsInterval = 100
purityRepetitions = 5
purityThreshold = 0.75
```

As described before, phonemic classes can either be fixed and immutable if `useFlexiblePhonology = FALSE` (in this case, `phoneme` must be set) or they can be computed using unsupervised learning algorithms (GMM and NMF) if `useFlexiblePhonology = TRUE`. In the latter case, three arguments are used to configure GMM and NMF. `computeGMMsInterval` takes a full positive number and defines after how many memorised tokens an agent recomputes their sub-phonemic classes. Be aware that this way of generating phonemic classes is computationally expensive and may lead to long computation times if the interval is set to a relatively low value.

Sub-phonemes must surpass the `purityThreshold`, i.e. the fraction of exemplars in a sub-phoneme belonging to a designated set of words, in order to be identified as a sub-phoneme. This threshold is a number between 0 and 1. Since there is an element of stochasticity to the NMF algorithm, the optimal number of sub-phonemes is determined by running NMF `purityRepetitions` times.

2.6 Interactions

```
interactionPartners = "betweenGroups"
speakerProb = NULL
listenerProb = NULL
```

The argument `interactionPartners` can be set to specify from which groups the two interacting agents shall come.

- `"random"`: It does not matter from which group an agent comes
- `"withinGroups"`: Speaker and listener must come from the same group
- `"betweenGroups"`: Speaker and listener must be members of different groups

The arguments `speakerProb` and `listenerProb` can be used to introduce an imbalance regarding the probability with which one or more agents are chosen to be speakers or listeners in an interaction. Both arguments take a vector of numbers (one number per agent). The numbers do not need to sum up to one, as they will be normalised internally. If left `NULL`, all agents will get equal chances to be selected as speakers or listeners in an interaction.

2.7 Runs

```
runs = 1
nrOfSnapshots = 10
interactionsPerSnapshot = 100
```

The model offers two ways of performing simulations: either as single runs or as multiple, parallel runs. Multiple runs of the same simulation can offer insights into the stability or robustness of the results. The argument `runs` specifies how many runs are computed, the default being a single run.

A simulation is a sequence of `nrOfSnapshots * interactionsPerSnapshot` interactions. At every `interactionsPerSnapshot` interactions, a snapshot of all agents' memories is saved.

2.8 Other options

```
rootLogDir = "./logDir"
notes = ""
```

The argument `rootLogDir` specifies the relative or absolute path to the logging directory where the simulation results will be saved. If the directory does not exist, it is created when running the simulation. Notes regarding the simulation can optionally be given in `notes`. This is useful because the arguments given to `run_simulation()` will be saved for each simulation, so the argument `notes` can help to specify the intention or purpose of running a specific simulation.

3 Analysing Results and Managing Simulations

3.1 Demo

There is a demo of this agent-based model available which uses the `u_fronting` dataset. The demo uses the default values for all arguments of `run_simulation()` apart from those that refer to the input data frame (see this section). The following function takes no arguments and starts the demo simulation:

```
run_demo_simulation()
```

3.2 Data Structures

Running a simulation causes multiple new directories and files to be created throughout the process. First of all, the root logging directory (as set by rootLogDir) is created if it did not previously exist. In this directory, all simulations will be saved under a name that consists of "ABM" and the date and exact time when the simulation was started. An example of that is `ABM20211102135337`: a simulation that was started on November 11th, 2021, at 1:53pm and 37 seconds. The root logging directory also contains the simulations register which is saved as an `.rds` and updated automatically whenever a new simulation is saved in the same directory. The next section is concerned with how to manipulate the simulations register manually.

In a simulation directory, there are two files and as many numerically named directories as there are runs. The two files are called `input.rds` and `params.yaml`. The first is the input data file as specified by inputDataFile, which, during the simulation, was loaded as a `data.table`. Its columns were renamed and the resulting data frame was saved as `input.rds`.

`rds` is a compressed file type specific to R which is used to save and restore single R objects. The R function to load `.rds` files is `readRDS()`. This function can be used, for example, to load

`input.rds`:

```
input.df <- readRDS("path/to/input.rds")
```

The second file in a simulation directory is `params.yaml`. YAML is a simple text format which can be opened with any text editor. `params.yaml` contains a list of the arguments including their values given to `run_simulation()`, as well as some more parameters that were created in the process of validating the arguments of `run_simulation()`. The function `get_params()` can be used to load the `params.yaml` file of a simulation by giving it the root logging directory and the simulation's name, for instance:

```
params <- get_params(rootLogDir = "./logDir",
                     simName = "ABM20211102135337")
```

The result of that function is a list.

`soundChangeR` also provides several wrappers for `readRDS()` in order to load the population, interactions log, cache, and input data file. The latter can be loaded using `load_input_data()` which takes the parameter list (as loaded by `get_params()`) as an argument.

```
input.df <- load_input_data(params = params)
```

The difference between `load_input_data()` and `readRDS(input.rds)` as shown above is that the latter loads the saved input file, while the former loads the original input data file, performs some conversions, and returns the loaded data as a `data.table`; in fact, `load_input_data()` is used during the simulation to load the input data file. The results of the two functions should be the same.

Further wrappers of `readRDS()` include `load_pop()`, `load_intLog()`, and `load_cache()` all of which take three arguments: `logDir` (the path to the simulation directory), `runs` (the runs to be loaded), and `snaps` (the snapshots to be loaded). In the subdirectories of the simulation directory, e.g. `ABM20211102135337/1` or `ABM20211102135337/2`, the snapshots of the agent population, the interactions log, and the cache are saved as `.rds` files. The files `pop.X.rds` (where `X` stands for a snapshot; `0 <= X <= nrOfSnapshots`) contain the agents' memories at the given time during the simulation as a data frame. These can be loaded using `load_pop()`, for example as follows:

```
pop <- load_pop(logDir = "./logDir/ABM20211102135337",
                runs = 1,
                snaps = c(0, 100))
```

Using these arguments, the function loads `./logDir/ABM20211102135337/1/pop.0.rds` and `./logDir/ABM20211102135337/1/pop.100.rds` and binds them together into a `data.table`. Both `runs` and `snaps` can either take a single full positive number or a vector of full positive numbers. The population data frame consists of the following columns:

- `run`: character column indicating the run
- `P1` etc.: all numeric columns starting with `P` are the acoustic features
- `word`: character column indicating the word class of the given exemplar

- `phoneme`: character column indicating the phonemic class of the given exemplar
- `nrOfTimesHeard`: numeric column indicating how often the corresponding agent has memorised tokens of the word class in `word`
- `producerID`: numeric column indicating the ID of the agent who produced the given exemplar
- `exemplar`: list of the feature columns
- `agentID`: numeric column indicating the ID of the agent in whose memory the given exemplar is stored
- `speaker`: character column indicating the agent's speaker code
- `group`: character column indicating the agent's group
- `snapshot`: character column indicating the snapshot

Given a data frame like the population or input data as an argument, the function `get_Pcols()` returns the names of the feature columns and `get_N_Pcols()` returns the amount of feature columns.

```
get_Pcols(data = pop)
get_N_Pcols(data = input.df)
```

The interactions log is saved in files called `intLog.X.rds` (where `X` again stands for snapshot; `1 <= X <= nrOfSnapshots`) and can be loaded using `load_intLog()`, for example:

```
intLog <- load_intLog(logDir = "./logDir/ABM20211102135337",
                      runs = 1,
                      snaps = c(1, 100))
```

In this case, the snapshots start at 1 (i.e. when the first nrOfInteractions interactions have taken place), not at 0 (i.e. before the first interactions). The interactions log contains the following columns:

- `run`: character column indicating the run
- `P1` etc.: all numeric columns starting with `P` are the acoustic features
- `snapshot`: character column indicating the snapshot
- `word`: character column indicating the word class of the given exemplar
- `producerID`: numeric column indicating the ID of the agent who produced the given exemplar
- `producerPhoneme`: character column indicating the phonemic class of the given exemplar according to the agent speaker
- `producerNrOfTimesHeard`: numeric column indicating how often the agent speaker has memorised tokens of the word class in `word`
- `perceiverID`: numeric column indicating the ID of the agent who perceived the given exemplar
- `perceiverPhoneme`: character column indicating the phonemic class of the given exemplar according to the agent listener
- `perceiverNrOfTimesHeard`: numeric column indicating how often the agent listener has memorised tokens of the word class in `word`
- `accepted`: logical column indicating whether the exemplar was memorised by the agent listener
- `rejectionCriterion`: character column indicating which memorisation criterion was responsible for rejecting the exemplar (i.e. if `accepted` is `FALSE`); if two memorisation

criteria were applied and both were failed, the first of them is given in this column

Finally, the cache is actually part of the agents' memories which is used for storing some agent-specific statistics and statistical models. The files are called `cache.X.rds` (where `X` again stands for snapshot; `0 <= X <= nrOfSnapshots`) and can be loaded into a `data.table` using `load_cache()`, e.g. as follows:

```
cache <- load_cache(logDir = "./logDir/ABM20211102135337",
                    runs = 1,
                    snaps = c(0, 100))
```

The cache contains the following columns:

- `run`: character column indicating the run
- `snapshot`: character column indicating the snapshot
- `agentID`: numeric column indicating the ID of the agent
- `name`: character column with one of `nFeatures` (number of acoustic features), `qda` (quadratic discriminant analysis for posterior probability memorisation criteria), `GMM` (Gaussian Mixture Model if `useFlexiblePhonology` is `TRUE`), `nAccepted` (total number of tokens that the agent has accepted and memorised), `nForgotten` (total number of exemplars that the agent has removed from memory)
- `value`: list of values corresponding to the cached object indicated by `name`
- `valid`: logical column indicating whether the cached object is valid, i.e. used at the given `snapshot` during the simulation

3.3 Analysis of Results

Three functions were implemented to produce some basic plots of the simulation results. These are not exhaustive, i.e. there are certainly more metrics that might be interesting depending on the simulation settings and input data. However, the following plotting functions are a start to understanding whether any change emerged from the simulation and if so, why.

A simulation was run on the `u_fronting` dataset, i.e. younger and older SSBE speakers interacted with one another, exchanging exemplars of 11 word classes in a three-dimensional DCT-based acoustic space. The settings of this simulation that differed from the default arguments of `run_simulation()` were:

```
SMOTENN = 10.0
useFlexiblePhonology = TRUE
runs = 5
nrOfSnapshots = 250
interactionsPerSnapshot = 1000
```

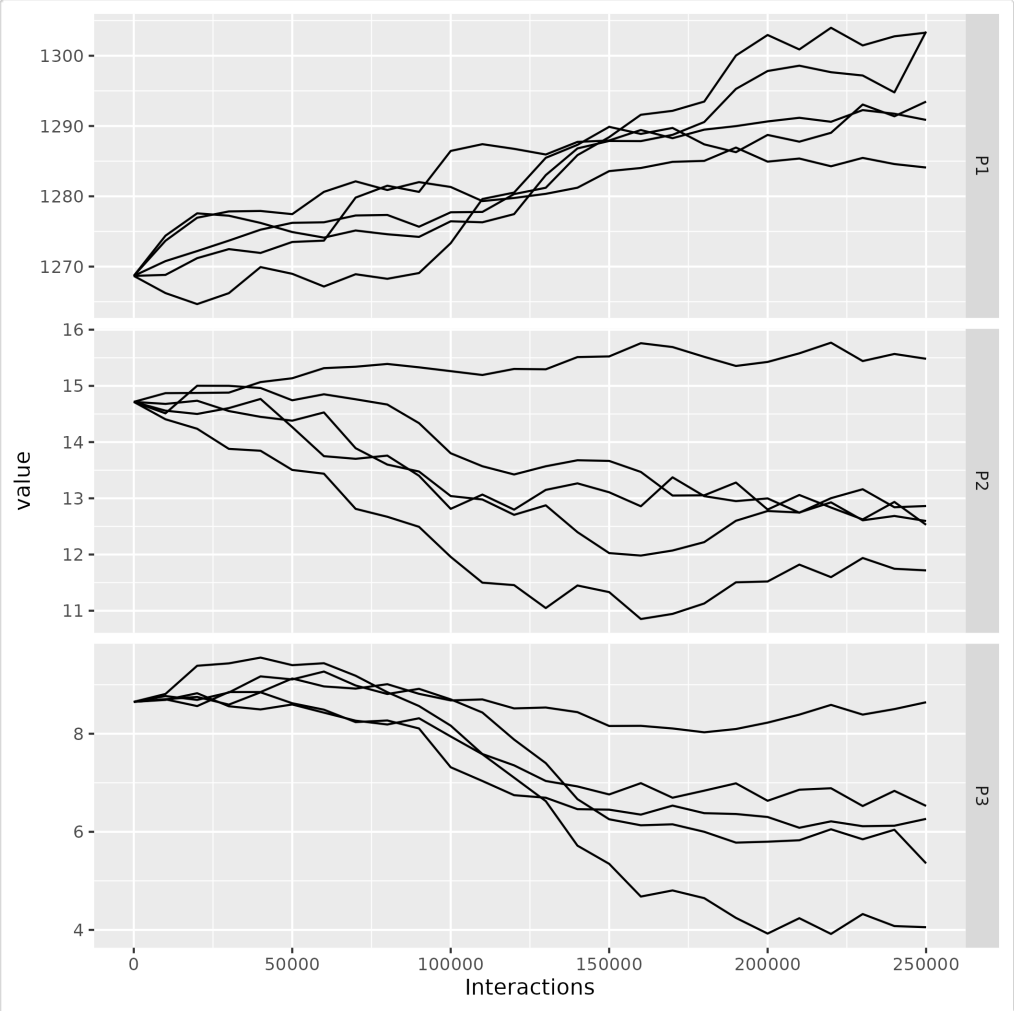
We use some of the functions described in the previous section to load the data from this simulation. The population data frame `pop` is altered to contain the canonical phonemes /i:, u:, ju:/. These canonical phonemes as well as the information on the agents' group membership is added from `pop` to the interactions log `intLog`.

```
rootLogDir <- "./logDir"
simulationName <- "ABM20211111100942"
logDir <- file.path(rootLogDir, simulationName)
params <- get_params(rootLogDir, simulationName)

pop <- load_pop(logDir,
                runs = 1:5,
                snaps = seq(0, 250, by = 10)) %>%
  mutate(
    canonical =
      case_when(word %in% c("seep", "heed", "keyed", "feed") ~ "i:",
                word %in% c("soup", "who'd", "cooed", "food") ~ "u:",
                word %in% c("hewed", "queued", "feud") ~ "ju:")
    )
Pcols <- get_Pcols(pop)
intLog <- load_intLog(logDir,
                      runs = 1:5,
                      snaps = seq(1, 250, by = 10)) %>%
  left_join(
    pop %>%
      select(speaker, group, word, canonical, agentID) %>%
      unique(),
    by = c("perceiverID"="agentID", "word")
  )
```

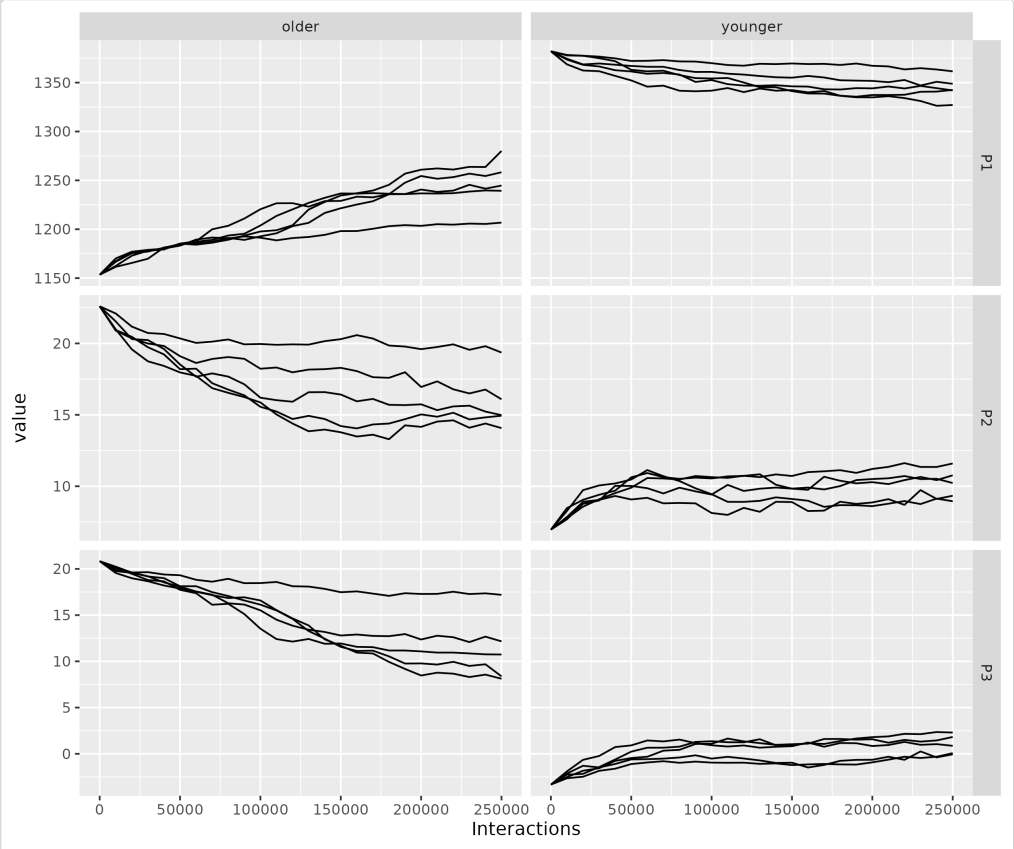
The first plotting function is called `plot_centroids()` and receives the population data frame `pop`, the columns with the acoustic features `Pcols`, grouping variables `groupVar`, and the parameter list `params`. The resulting plot shows the mean acoustic features over simulation time, i.e. over the course of the interactions. The default grouping variables are `snapshot` and `run`, of which `snapshot` is obligatory. If `run` is part of `groupVar`, it is always plotted as "group" in the aesthetic mappings, so that there is one line per run in each panel. Further grouping variables are either colour-coded or plotted in facets. More than four grouping variables cannot be plotted.

```
plot_centroids(pop, Pcols, groupVar = c("snapshot", "run"), params)
```



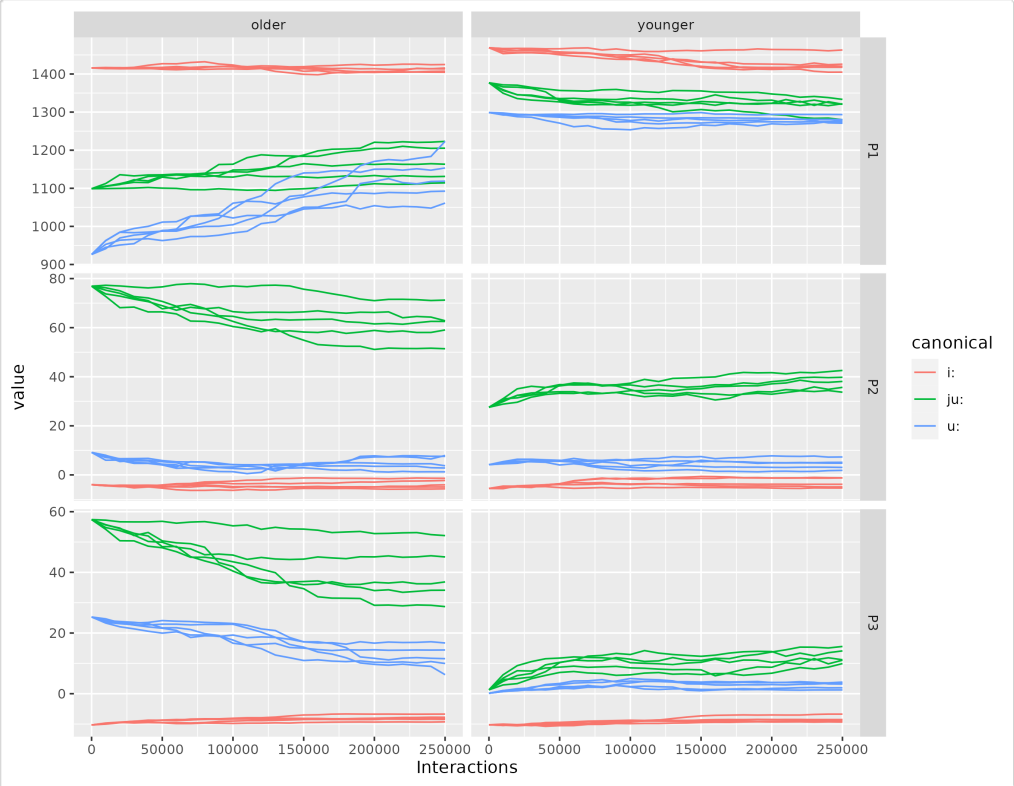
This plot shows that P1 (i.e. DCT0) increased over simulation time when aggregated across the population and canonical phonemes in all five runs, and P2 and P3 (i.e. DCT1 and DCT2) decreased slightly in four out of five runs. Since there were two agent groups with markedly different starting points, let's differentiate the plot also by group:

```
plot_centroids(pop,
               Pcols,
               groupVar = c("snapshot", "run", "group"),
               params)
```



This shows that older agents changed more than younger ones – an important finding, given that in this case we would expect older agents to adapt to younger agents' vowel variants, rather than vice versa. Since /i:/ and /u:/ are still aggregated in the plot, the following plot colour-codes by the canonical phonemes.

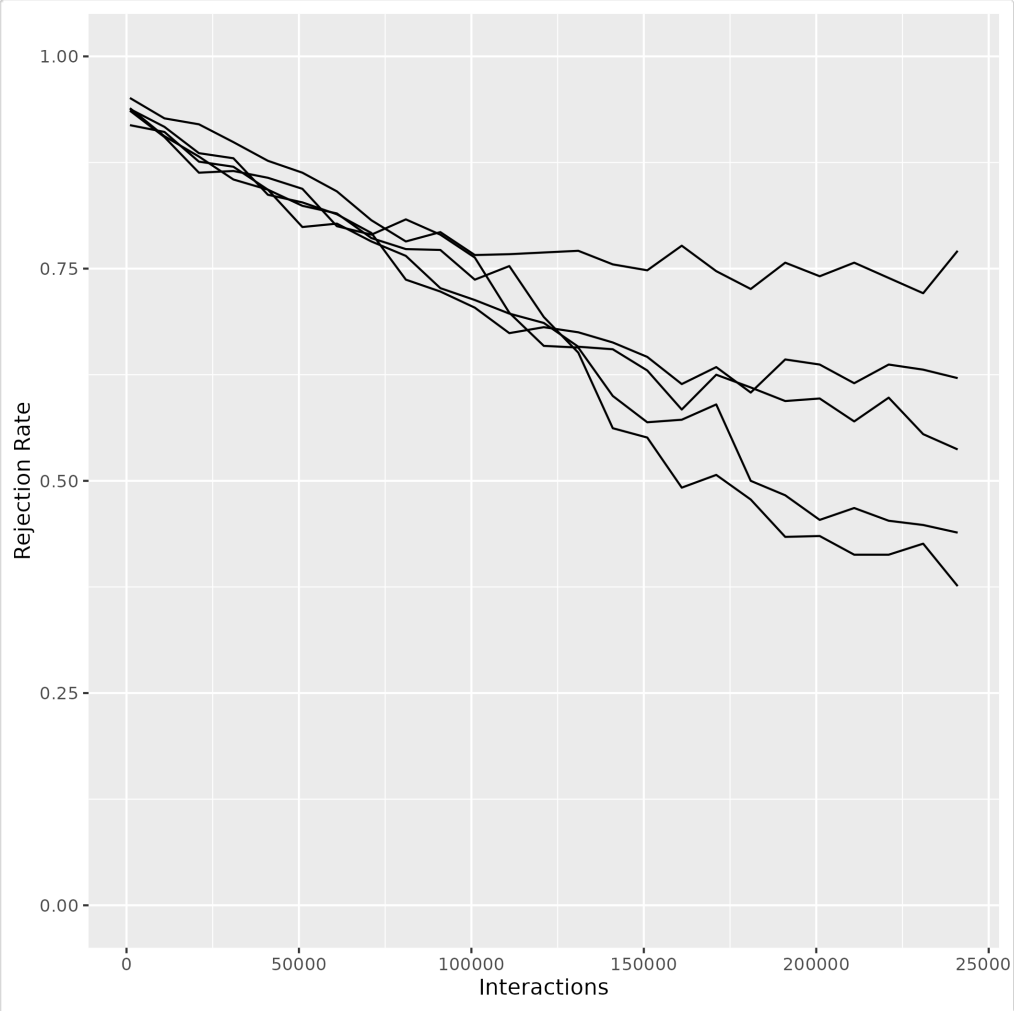
```
plot_centroids(pop,
               Pcols,
               groupVar = c("snapshot", "run", "group", "canonical"),
               params)
```



This plot shows that older agents' /u:/ shifts towards that of younger agents, especially in P1 (DCT0) and in all five runs. The same holds for /ju:/, although to a lesser degree. Younger agents' vowels and older agents' /i:/ does not change and therefore shows that change is not an inevitable outcome of the simulations.

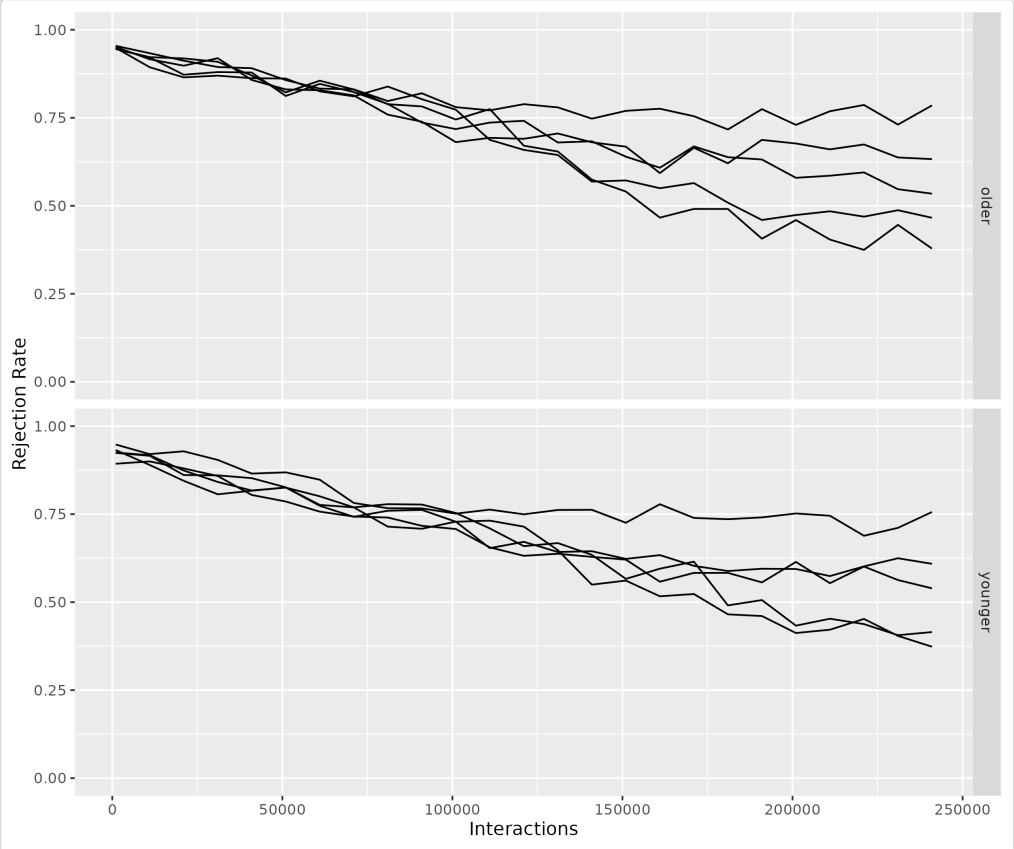
The second plotting function, called `plot_rejection()`, shows the rejection rate over simulation time. Rejection rate is calculated as \(1-\frac{nr.~of~accepted}{nr.~of~perceived}\). The function takes the interactions log `intLog`, grouping variables `groupVar`, and the parameter list `params` as arguments. The constraints on the grouping variables are the same for `plot_rejection()` as they are for `plot_centroids()`.

```
plot_rejection(intLog, groupVar = c("snapshot", "run"), params)
```



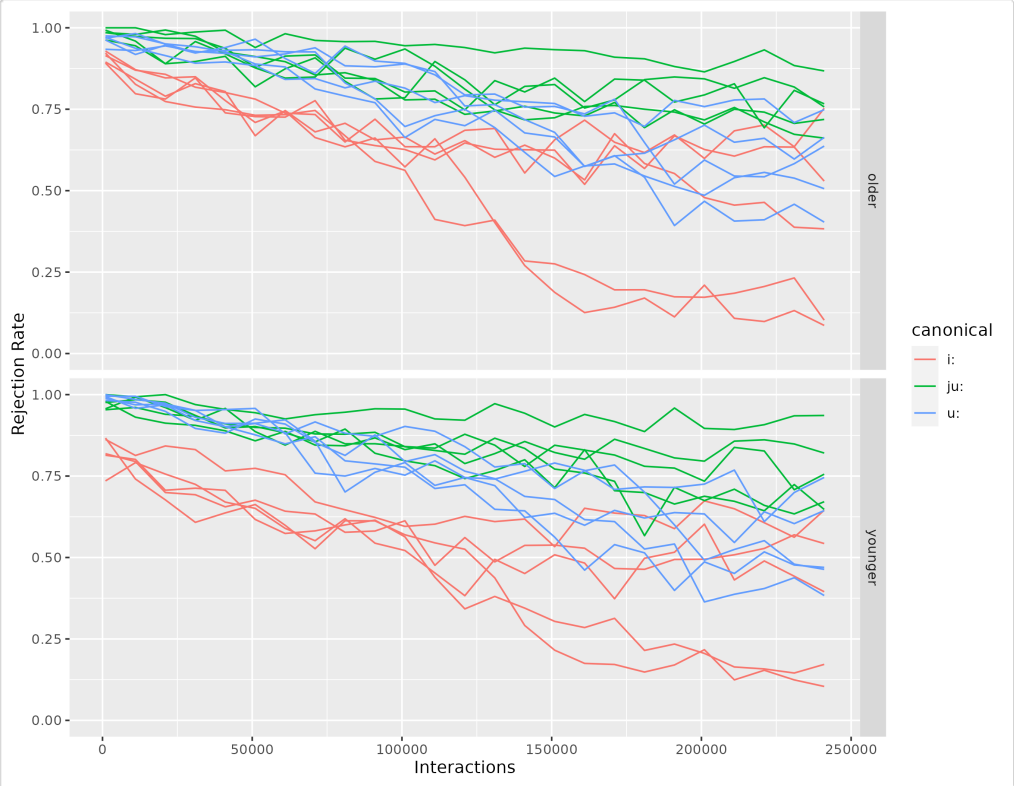
The rejection rate can range between 0 and 1. In this plot, it can be seen that the rejection rate decreases over simulation time in all five runs. This means that the agents adapt to each other since they start to accept more and more tokens. A separation by agent group might be interesting, but in this case, there does not seem to be a big difference in rejection rate between the two:

```
plot_rejection(intLog, groupVar = c("snapshot", "run", "group"), params)
```



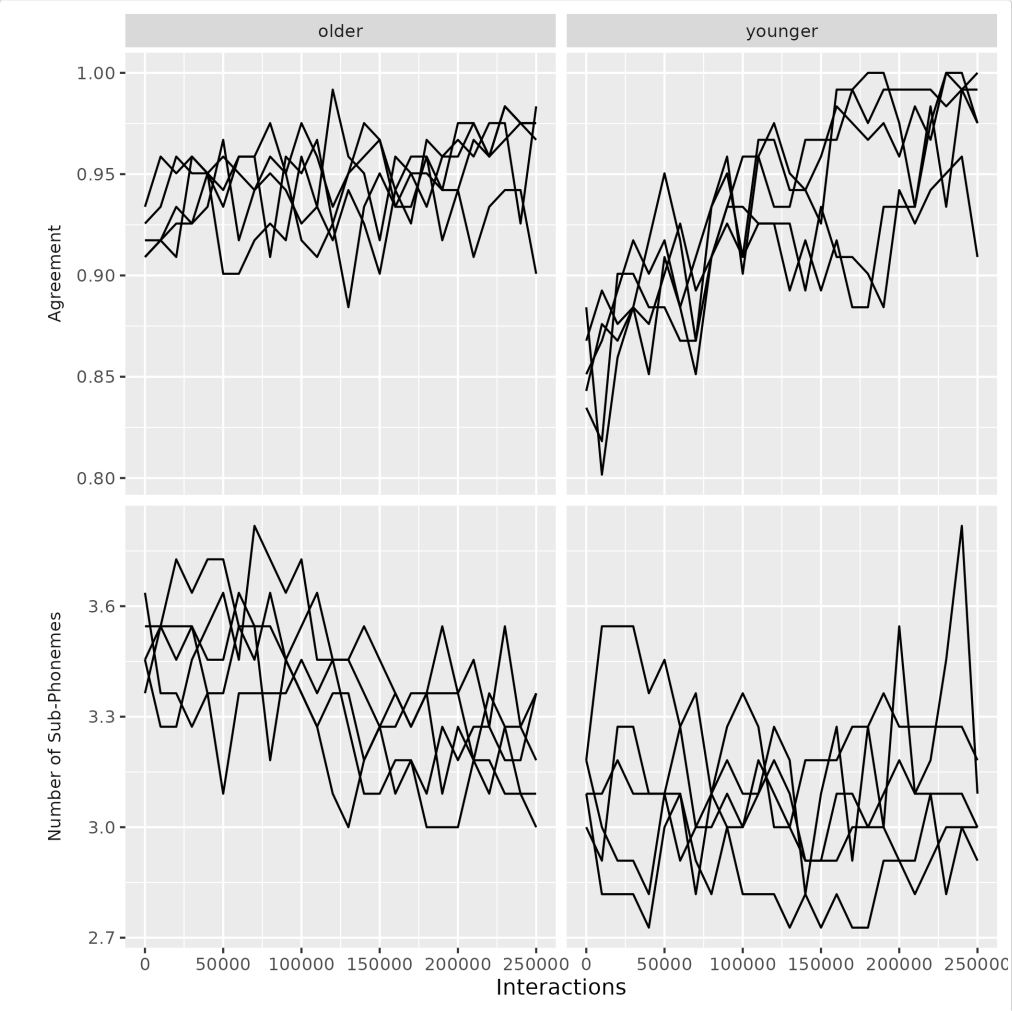
When we differentiate the same plot also by canonical phoneme, it can be seen that the rejection rate is lower for /i:/ than for /u:, ju:/. This is because the agent groups are very similar in how they pronounce /i:/, so they mostly accept each others exemplars.

```
plot_rejection(intLog,
               groupVar = c("snapshot", "run", "group", "canonical"),
               params)
```



Finally, the function `plot_phonology()` creates a plot of two metrics that need to be interpreted together and only make sense when `useFlexiblePhonology = TRUE`. The first metric is the number of sub-phonemic classes over simulation time. The second is a measurement of the agreement of the sub-phonemic classes with the canonical phonemes. In order to compute this measure, the user must add the canonical phonemes to the `pop` data frame as shown above and give the column's name to `plot_phonology()` as argument `canonical`. The other two arguments to the function are `pop` (the population data frame) and `params` (the parameter list). If there were agent groups, they are automatically plotted in different columns. Runs are grouped per panel, as shown below for the five runs of this simulation.

```
plot_phonology(pop, canonical = "canonical", params)
```



Using the `purity()` function from the `NMF` package, the agreement between sub-phonemes and canonical phonemes is calculated as $\frac{1}{N_w} \cdot \sum_p N_{\textrm{majority}}(p)$. N_w is the total number of word classes (e.g. 11 in case of the `u_fronting` data frame) and $N_{\textrm{majority}}(p)$ is the number of unique word classes in a sub-phoneme p that belong to the majority canonical phoneme, i.e. the canonical phoneme which is associated with the largest number of unique word classes in p. The `u_fronting` data frame contains 4 /i/-words, 4 /u/-words, and 3 /ju/-words. So if an agent had three sub-phonemes, $p1$ containing exemplars of the 4 /i/-words, $p2$ containing exemplars of the 3 /ju/-words as well as 1 /u/-word, and $p3$ containing exemplars of the remaining 3 /u/-words, the agreement is $\frac{4 + 3 + 3}{11} \approx 0.91$. If an agent had two sub-phonemes, $p1$ containing exemplars of 3 /i/-words, 2 /ju/-words, and 1 /u/-word and $p2$ containing exemplars of the remaining 1 /i/-word, 1 /ju/-word/, and 3 /u/-words, the agreement is $\frac{3 + 3}{11} \approx 0.55$.

It is important to take into account both the number of sub-phonemes and the agreement, because on their own you may draw the wrong conclusions. For instance, if many sub-phonemes (e.g. 10 or so) have developed over the course of the simulation, it may look like a phonological split has occured; however, if the overall agreement is low this actually indicates

that the sub-phonemes contain a balanced mixutre of exemplars from all canonical phonemes, which might rather point to a merger. On the other hand, if all sub-phonemes contain exemplars of only one word class each, the agreement is 1; so looking only at the agreement would indicate that the sub-phonemes are in perfect agreement with the canonical phonemes when actually this is due to the over-formation of sub-phonemes.

For the exemplary simulation on the `u_fronting` simulation, the plot above shows the agreement and number of sub-phonemes over simulation time for older and younger speakers. From the scale of the top row, it can be seen that the overall agreement between sub-phonemes and canonical /i, u, ju/ is higher than 0.8 throughout the simulation. For older agents, the agreement is even higher than 0.9 and remains stable, for younger agents it increases over time. The number of sub-phonemes circle around 3, with older agents starting on average with a little more sub-phonemes than younger agents and then decreasing their number of sub-phonemes. From this plot, we can conclude that both agent groups' phonology becomes more similar to the canonical phoneme separation into /i/, /u/, and /ju/: older agents' sub-phonemes separate /i/-, /u/-, and /ju/-words nicely from the start and over the course of the interactions they additionally decrease the number of sub-phonemes to approx. 3; younger agents have the expected three sub-phonemes which, at simulation start, are not exactly in agreement with the canonical phonemes, but come to be over time.

3.4 Managing Simulations

There are four functions in `soundChangeR` which help to manage simulations that are listed in `simulations_register.rds`. The first is `filter_simulations()` which goes through the registered simulation and returns the simulation names of all simulations that match the given filters. These filters must be arguments from `run_simulation()` and their desired values. Here are three examples:

```
filter_simulations(rootLogDir = "./logDir",
                   forgettingRate == 1)
filter_simulations(rootLogDir = "./logDir",
                   forgettingRate == 1 & useFlexiblePhonology == TRUE)
filter_simulations(rootLogDir = "./logDir",
                   condList = list(
                     forgettingRate = 1,
                     useFlexiblePhonology = TRUE
                   ))
```

So `filter_simulations()` needs the root logging directory as the first argument, and then either a single filtering argument using logical operators, or a list of filtering arguments without logical operators.

The second helper function is `delete_simulation()`. This function needs the root logging directory and a simulation name as arguments and removes the simulation from the simulations register (but does not delete the simulation results). In the following example, the simulation `./logDir/ABM20211102135337` is removed from the simulations register.

```
delete_simulation(rootLogDir = "./logDir",
                  simName = "ABM20211102135337")
```

The next two functions remove simulations from the register as well as permanently deleting the simulation's results. `purge_simulation()` takes the same arguments as

`delete_simulation()`, e.g.:

```
purge_simulation(rootLogDir = "./logDir",
                 simName = "ABM20211102135337")
```

The function `purge_uncompleted_simulations()` only takes the root logging directory as an argument, then searches for incomplete simulations in the register and deletes those and their results, e.g.:

```
purge_uncompleted_simulations(rootLogDir = "./logDir")
```

4 Recommended Literature

Gubian, M., Cronenberg, J., and Harrington, J. (under review): Phonetic and Phonological Sound Changes in an Agent-Based Model. Speech Communication.

Cronenberg, J. (in prep.): New Approaches to the Study of Sound Change: The Case of Aspiration in Andalusian Spanish. Dissertation, LMU Munich, chapter 3.

Harrington, J., Kleber, F., Reubold, U., Schiel, F., and Stevens, M. (2018): Linking Cognitive and Social Aspects of Sound Change Using Agent-Based Modeling. Topics in Cognitive Science, pp. 1-22.

Harrington, J., and Schiel, F. (2017): /u/-fronting and agent-based modeling: The relationship between the origin and spread of sound change. Language 93 (2), pp. 414-445.

B.4 Calculating Sub-Phonemic Classes

This section provides mathematical details as well as an in-depth example of Gaussian Mixture Models and non-negative matrix factorisation which are used to calculate sub-phonemic classes in the agent-based model. This appendix is supplementary to section 3.2.5.

B.4.1 Gaussian Mixture Models-based Clustering and Classification

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities (Reynolds, 2009). Its general form is:

$$f(x|\lambda) = \sum_{j=1}^{G} w_j \cdot g(x|\mu_j, \Sigma_j) \tag{B.1}$$

where x is a continuous-valued d-dimensional vector, G is the number of components, $g(\cdot)$ is the multidimensional Gaussian density function and $\lambda = \{w_j, \mu_j, \Sigma_j\}$, $j = 1, \ldots, G$ indicates the set of weights w_j satisfying $\sum_{j=1}^{G} w_j = 1$, the d-dimensional mean vectors μ_j and the $d \times d$-dimensional covariance matrices Σ_j.

Given a set of N d-dimensional data vectors $\{x_i\}$, $i = 1, \ldots, N$, and fixing the number of components G, a model like Eq. (B.1) can be obtained by maximum likelihood (ML) estimation, usually applying the expectation-maximisation (EM) algorithm, as there is no closed-form solution for the maximisation of Eq. (B.1) with respect to λ. Often constraints are applied on the structure of Σ_j, e.g. diagonal, or parameters are tied across components, e.g. $\Sigma_j = \Sigma, \forall j$. The number of components can be estimated by first estimating several candidate models like Eq. (B.1), each one with a different number of components G, and then applying the Bayesian information criterion (BIC), or alternative IC methods, to select the best model.

Eq. (B.1) can be used as a clustering model for the data vectors $\{x_i\}$ that were used to estimate its parameters. By interpreting the mixture components as clusters, any vector x, from the training dataset or otherwise, can be assigned to a cluster j by applying the maximum *a posteriori* (MAP) criterion:

$$\begin{aligned} Cluster(x) &= \arg\max_j \frac{w_j \cdot g(x|\mu_j, \Sigma_j)}{\sum_{k=1}^{G} w_k \cdot g(x|\mu_k, \Sigma_k)} \\ &= \arg\max_j w_j \cdot g(x|\mu_j, \Sigma_j) \end{aligned} \tag{B.2}$$

When information on class membership of the data is available, i.e. the vectors x_i are paired with corresponding class labels $y_i \in \{1, \ldots, K\}$, where K is the number of classes, it is possible to use the form Eq. (B.1) as building block for a hierarchical classification model (Mixture Discriminant Analysis, MDA, Fraley & Raftery, 2002). MDA is a generalisation of Linear and Quadratic Discriminant Analysis (LDA, QDA) that allows the density of each class to be of the form of Eq. (B.1). Class membership y for a new vector x is obtained by applying Bayes's rule:

$$\begin{aligned} Class(x) &= \arg\max_y Pr(x \in y) \\ &= \arg\max_y \frac{\tau_y \cdot f_y(x|\lambda_y, G_y)}{\sum_{k=1}^{K} \tau_k \cdot f_k(x|\lambda_k, G_k)} \\ &= \arg\max_y \tau_y \cdot f_y(x|\lambda_y, G_y) \end{aligned} \tag{B.3}$$

where $\{\tau_y\}, y \in \{1, \ldots, K\}$ are the proportions of members of class y in the training set, $f_y(\cdot)$ are functions of the form of Eq. (B.1), each having in general a different parameter set λ_y and number of components G_y.

In `soundChangeR`, both GMM clustering and MDA classification are implemented using the R package *mclust* (Scrucca et al., 2016). The number of acoustic components, as well as a suitable set of constraints on the parameter set λ, are determined by the BIC criterion as implemented by the function

`Mclust`. The MDA model is estimated using function `MclustDA`, where the number of components of each phonemic class is fixed according to the mapping obtained from the NMF procedure (see Appendix B.4.3). For example, if clusters a_1 and a_2 map to sub-phoneme p_1 and a_3 maps to p_2, then the number of components for p_1 and p_2 are fixed to 2 and 1, respectively. This helps to preserve the original purely acoustic-based structure expressed by the GMM clustering model.

B.4.2 Non-negative Matrix Factorisation

Non-negative Matrix Factorisation (NMF, D. D. Lee & Seung, 2001) is a family of algorithms that given a $m \times n$ input matrix C with non-negative entries determines an approximation of it in the form of a product of two matrices, W and H, also with non-negative entries. Formally:

$$C \approx W \cdot H \tag{B.4}$$

where W has dimension $m \times r$, H has dimension $r \times n$, and usually $r < \min(m, n)$. Once r is fixed, Eq. (B.4) is solved by minimising the difference between C and its approximation $W \cdot H$, i.e. W and H are the solution to:

$$\min_{W,H} \| C - W \cdot H \|_F^2, \text{ such that } W, H \geq 0, \tag{B.5}$$

where $\| \cdot \|_F$ is the Frobenius norm, the analogous of Euclidean norm for matrices.

In applications, each of the n columns c_j of C represents an observation or feature vector, where the m elements of c_j represent counts, frequencies, energies or other non-negative quantities. For example, in text mining the n columns of C represent documents and the m rows word counts or frequencies (Shahnaz et al., 2006). As the number of words × documents can be in the thousands × millions, it is convenient to find a compact approximation of C in the form of Eq. (B.4) with r as small as possible, say in the tens. Similarly to PCA, the columns of W can be seen as bases or principal components capturing the fundamental traits of the columns of C (though the columns of W are not

constrained to be orthogonal), while each column h_j of H determines the linear combination of columns of W that best reconstructs a corresponding input column c_j. Differently from PCA, the non-negativity of W and H in Eq. (B.4) allows to interpret the composition of columns of W as strictly additive contribution from each basis in the quantities determined by the positive coefficients in H, where no contribution can undo (i.e. subtract) parts of the others. For example, in text mining the r columns of W are interpreted as topics, and each document c_j, codified by the number of occurrences of m selected words (i.e. bag of words representation), is a weighted sum of topics, the weights being the elements of column h_j of H.

A further step in the interpretation of Eq. (B.4) is achieved whenever W and H are sufficiently sparse, in which case the columns of H tend to ideally have only one large entry, which identifies only one column of W. This means that each of the n input column vectors of C can be approximated by just a coefficient multiplying one of the $r < n$ columns w_k of W, i.e. $c_j \approx H(k,j) \cdot w_k$ when $H(k,j) \gg H(i,j)$, $\forall i \neq k$. The particular column w_k approximating a given input vector c_j can be interpreted as its centroid or prototype, analogously to other prototype-based clustering algorithms, e.g. k-means (see Ding et al., 2005, for a formal statement). This interpretation allows to utilise NMF as a clustering algorithm, with r being the number of clusters. In order to improve the performance of NMF as clustering method, several sparse solutions have been proposed (e.g. J. Kim & Park, 2008; Pascual-Montano et al., 2006). As for most clustering algorithms, the optimal number of clusters r has to be determined by applying a model selection criterion external to the algorithm itself.

In `soundChangeR`, NMF clustering is implemented using the the R package *NMF* (Gaujoux & Seoighe, 2010). In particular, the NMF algorithm version by Pascual-Montano et al. (2006) is selected, which produces sparser solutions (setting method option to 'nsNMF' in the `nmf` command). The cluster assignment is determined by taking the arg max of each column of matrix H.

B.4.3 Derivation of Acoustic and Sub-Phonemic Representations

Table B.1 explains the symbols used in this appendix as well as in Appendices B.5.2 and B.6.

B.4.3.1 Overview

Each agent periodically re-estimates its acoustic and sub-phonemic models following a two-stage process schematised in Figure B.1. First, a GMM clustering model as in Eq. (B.1) is estimated from the acoustic representation of all the $N = N_e$ exemplars stored in memory, i.e. the set $\{x(e_i)\}$, $i \in \{1,\dots,N_e\}$, disregarding any information on sub-phonemic classes established in previous estimations. The number of acoustic clusters $N_a = G$ is estimated by BIC. By applying Eq. (B.2) each exemplar is assigned to an acoustic cluster $a \in \{1,\dots,N_a\}$ corresponding to a single Gaussian component. The acoustic cluster membership information associated to each exemplar $\{a(e_i)\}$, but *not* the corresponding acoustic representation $\{x(e_i)\}$, is fed to the NMF-based procedure, which combines it with the word membership $\{w(e_i)\}$ to determine which acoustic clusters shall be joined together, i.e. a mapping from the acoustic labels $\{1,\dots,N_a\}$ to the sub-phonemic labels $\{1,\dots,N_p\}$, $N_p \leq N_a$. This mapping allows to indirectly associate each exemplar with a sub-phonemic class, i.e. $p(a(e_i))$. Using these class assignments, an MDA classification model is estimated, where the number of classes coincides with the number of sub-phonemic classes estimated by NMF, i.e. $K = N_p$ in Eq. (B.3). NMF operates on a matrix where each cell contains the exemplar count of a given word (row) in a given acoustic cluster (column). The result is a grouping of columns having similar count patterns, which are then merged together. Sub-phonemic classes are represented by those columns, while each word is assigned to a class by picking the column with the highest count for that word.

The two-stage process described above is executed periodically and independently by each agent. Within the same agent, each execution bears no memory of the acoustic and phonological representations derived in previous

Symbols	Meaning
x	an acoustic vector, e.g. $[1.2, -0.4, 0.88]^T$, when the acoustic space is 3-dimensional
d	number of dimensions of the acoustic space
w	a word type and a specific position within the word, e.g. *food*
e	an exemplar, i.e. a tuple (x, w)
$x(e)$	acoustic vector of exemplar e
$w(e)$	word type of exemplar e
N_e	number of exemplars in an agent's memory
N_a	number of acoustic clusters
N_p	number of sub-phonemic classes
N_w	number of word types
a	an acoustic cluster, determined by probabilistic membership to a Gaussian mixture component
p	a sub-phonemic class, determined by applying NMF to acoustic clusters
C_a	$N_w \times N_a$ matrix of exemplar counts, $C_a(j,k)$ is the count of exemplars of word w_j that belong to acoustic cluster a_k
C_p	$N_w \times N_p$ matrix of exemplar counts, $C_p(j,k)$ is the count of exemplars of word w_j that belong to sub-phonemic class p_k; these counts allow impurities. The sub-phonemic class of word type w_j is determined by picking $\arg\max_k C_p(j,k)$
$a(x(e)) = a(e)$	acoustic cluster associated to an acoustic vector
$p(a(e)) = p(a)$	sub-phonemic class associated to an acoustic cluster
$p(w(e)) = p(w)$	sub-phonemic class associated to a (position within a) word
Relations	**Comment**
$1 \leq N_p \leq N_a$	highest number of sub-phonemic classes occurs when each acoustic cluster constitutes a phonological class
$p(a(e)) \neq p(w(e))$	if an exemplar is 'impure', the sub-phonemic class associated to its location in the acoustic space is different from that associated to the word it belongs to

Table B.1: Symbols used in Appendices B.4.3, B.5.2, and B.6.

.

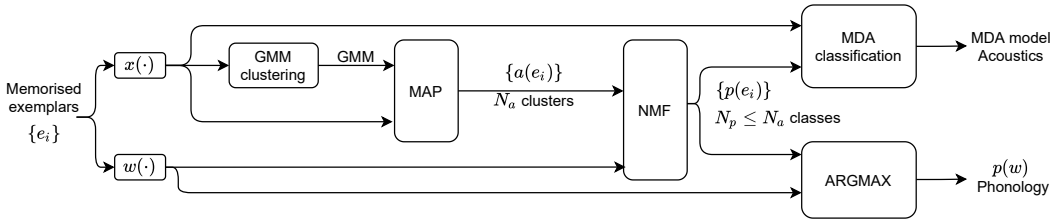


Figure B.1: General scheme of the implementation of acoustic and sub-phonemic representations in `soundChangeR`.

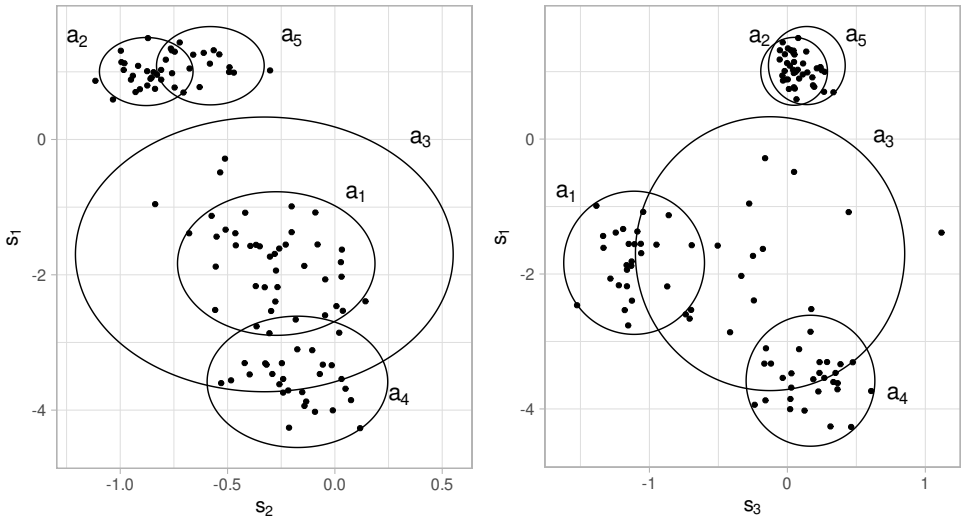
executions, a choice driven by implementation convenience. Continuity is indirectly guaranteed by making sure that only a small fraction of memory has changed (through exemplar memorisation and forgetting) between one execution and the next.

B.4.3.2 Identification of Sub-Phonemic Classes by NMF

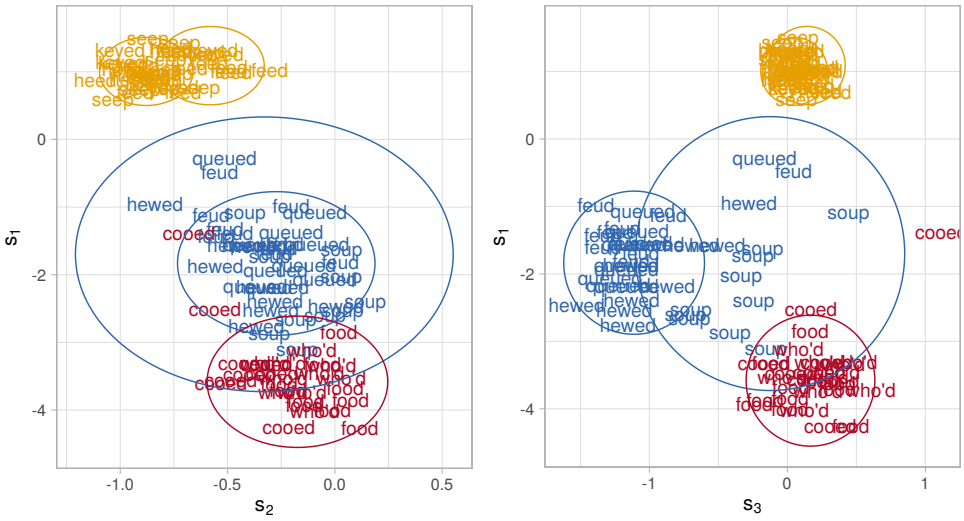
The input to NMF are (i) the set of acoustic cluster memberships $\{a(e_i)\}$, $i \in \{1,\ldots,N_e\}$, obtained by applying MAP as in Eq. (B.2) to the acoustic exemplars $\{x(e_i)\}$, and (ii) the set of word memberships of each exemplar, $\{w(e_i)\}$. These two pieces of information are combined into a $N_w \times N_a$ matrix of counts C_a, whose (j,k) element is the count of exemplars of word w_j that belong to acoustic cluster a_k, i.e. $C_a(j,k) = |\{i : w(e_i) = w_j, a(e_i) = a_k\}|$. This matrix is decomposed into the product of two matrices using NMF as in Eq. (B.4), i.e. $C_a \approx W \cdot H$, where the rank of the approximation corresponds to the number of sub-phonemic classes ($r = N_p$) and it is empirically determined as explained later on. Matrix H is further processed to obtain an indicator matrix $\widetilde{H}$, where each column h_k of H is substituted with a column of zeros and ones, with a single 1 at the position corresponding to the maximum element of h_k and zeros elsewhere. This is then used to recompute word counts according to the new N_p sub-phonemic clusters:

$$C_p = C_a \cdot \widetilde{H}^T \tag{B.6}$$

where C_p is a $N_w \times N_p$ count matrix and ^{T} is matrix transposition.



(a)



(b)

Figure B.2: Example of acoustic clusters (B.2a) and sub-phonemic classes (B.2b) in an agent. In B.2a, dots represent memorised exemplars in the acoustic space, and ellipses $a_1 - a_5$ identify acoustic clusters. In B.2b, the same exemplars are labelled according to the word they are associated with, while colours show their sub-phonemic classes. The acoustic space is three-dimensional (s_1, s_2, s_3) based on FPCA scores (cf. section 2.2.1.3); left and right views are projections on the (s_1, s_2) and (s_1, s_3) plane, respectively.

An example was taken from the Standard Southern British English dataset that is part of `soundChangeR` (Harrington et al., 2008; Harrington & Schiel, 2017, also see example in section 3.2.5, Figure 3.3). One of the 22 agents at simulation start (i.e. before any interaction has taken place) has the acoustic cluster counts C_a shown in Table B.2a, where we see that this agent divides its acoustic space into $N_a = 5$ clusters, a_1 to a_5, corresponding to the five Gaussian mixture components plotted in Figure B.2a. Applying Eq. (B.4) to C_a, where we fix the rank to $r = N_p = 3$, we obtain:

$$\underbrace{\begin{bmatrix} 0 & 0 & 2 & 7 & 0 \\ 0 & 3 & 0 & 0 & 7 \\ 9 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 10 & 0 \\ 0 & 9 & 0 & 0 & 1 \\ 8 & 0 & 2 & 0 & 0 \\ 0 & 7 & 0 & 0 & 3 \\ 9 & 0 & 1 & 0 & 0 \\ 0 & 9 & 0 & 0 & 1 \\ 3 & 0 & 6 & 1 & 0 \\ 0 & 0 & 0 & 10 & 0 \end{bmatrix}}_{C_a} \approx \underbrace{\begin{bmatrix} 0 & 0 & 0.31 \\ 0 & 0.25 & 0 \\ 0.25 & 0 & 0 \\ 0 & 0 & 0.34 \\ 0 & 0.25 & 0 \\ 0.25 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0.25 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0.25 & 0 & 0 \\ 0 & 0 & 0.34 \end{bmatrix}}_{W} \times \underbrace{\begin{bmatrix} 29 & 0 & 12 & 0 & 0 \\ 0 & 28 & 0 & 0 & 12 \\ 0 & 0 & 0 & 28 & 0 \end{bmatrix}}_{H} \tag{B.7}$$

where we note that the NMF solution is indeed sparse. For example, the first column of C_a is approximated by the first column of W multiplied by 29, the second column of C_a by the second column of W multiplied by 28, etc., each column of C_a is approximated by just one column of W.[16] This provides the foundation for using the NMF expansion as a clustering criterion, namely columns of C_a approximated by the same column of W, such as columns 1 and 3, belong together because they are approximately proportional, which in our application means that they roughly contain the same word types. By observing C_a in Table B.2a we can see that this is the case for acoustic clusters

[16] The reader unfamiliar with matrix multiplication may consult any text on linear algebra.

a_1 and a_3, which makes them good candidates for a merge. The merge of acoustic categories is obtained by constructing the indicator matrix $\widetilde{H}$ from H:

$$\widetilde{H} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

the rows of which indicate with 1s the columns of C_a which should be merged (summed) into columns of C_p, e.g. the first row indicates that the first and third column of C_a will be summed together to become the first column of C_p. This operation is formalised by Eq. (B.6), the result of which is Table B.2b.

The result in Eq. (B.6) defines the mapping from acoustic clusters to sub-phonemic clusters $p(a(e_i))$, which is not yet a definition based on word types, but rather on exemplars. For instance, according to Table B.2b two exemplars of *cooed* belong to sub-phonemic class p_1 and seven to p_3. To obtain a many-to-one mapping from word types (and position within word) to sub-phonemic classes we assign each word w to the sub-phonemic cluster with highest exemplar count, i.e.

$$p(w) = \arg\max_p C_p(w, p) \tag{B.8}$$

For example, $p(\text{cooed}) = p_3$, $p(\text{feed}) = p_2$, etc., as indicated by boldface figures in Table B.2b, by labels in Table B.2c, and by colours in Figure B.2b. We call the entities defined by the mapping in Eq. (B.8) sub-phonemic *categories* or *classes*, to distinguish them from the sub-phonemic *clusters* identified by exemplar counts in C_p. This distinction is crucial as it brings along the concept of cluster *purity*, which plays a central role in determining the optimal number of sub-phonemic clusters, as explained below.

Purity is defined for each cluster as the fraction of exemplars that belong to that cluster, according to the majority rule defined in Eq. (B.8), divided by its exemplar count. Purity of the whole clustering solution is defined as the minimum purity value among the clusters (other definitions are possible, e.g. taking the average instead of the minimum, cf. Manning et al. (2008)). Formally:

	Acoustic clusters				
Word	a_1	a_2	a_3	a_4	a_5
cooed	0	0	2	7	0
feed	0	3	0	0	7
feud	9	0	1	0	0
food	0	0	0	10	0
heed	0	9	0	0	1
hewed	8	0	2	0	0
keyed	0	7	0	0	3
queued	9	0	1	0	0
seep	0	9	0	0	1
soup	3	0	6	1	0
who'd	0	0	0	10	0

(a) C_a

	Sub-phonemic clusters		
Word	p_1	p_2	p_3
cooed	2	0	**7**
feed	0	**10**	0
feud	**10**	0	0
food	0	0	**10**
heed	0	**10**	0
hewed	**10**	0	0
keyed	0	**10**	0
queued	**10**	0	0
seep	0	**10**	0
soup	**9**	0	1
who'd	0	0	**10**

(b) C_p

	Sub-phonemic classes
Word	
cooed	p_3
feed	p_2
feud	p_1
food	p_3
heed	p_2
hewed	p_1
keyed	p_2
queued	p_1
seep	p_2
soup	p_1
who'd	p_3

(c) $p(w)$

Table B.2: Exemplar counts by word type (rows) and by acoustic cluster (columns, B.2a) or by sub-phonemic cluster (columns, B.2b). This is a snapshot at simulation start of the memory of one of the agents from the /u/-fronting dataset described in Harrington and Schiel (2017). In B.2b, boldface figures indicate which sub-phonemic cluster contains the majority of exemplars for each word type, which determines the assignment of words to sub-phonemic classes $p(w)$ in B.2c.

$$Purity = \min_k \left(\frac{|\{i : p(a(e_i)) = p(w(e_i)) = p_k\}|}{|\{i : p(a(e_i)) = p_k\}|} \right) \tag{B.9}$$

To illustrate, the purity of each cluster in Table B.2b is the sum of boldface counts divided by all counts in each column, e.g. $Purity(p_1) = \frac{39}{41} = 0.951$, which is the smallest value among the three ($Purity(p_2) = 1.0$ and $Purity(p_3) = \frac{27}{28} = 0.964$), so also the purity of the whole clustering solution according to Eq. (B.9). The optimal number of sub-phonemic clusters N_p is determined empirically by running several instances of NMF for each value of N_p from 2 to N_a, and choose the result that on average reaches the best trade-off between a high number of clusters and a high purity. This criterion in general disfavours the two extreme solutions, i.e. $N_p = 1$ and $N_p = N_a$. On the one hand, the highest number of clusters $N_p = N_a$ is probably not the solution with the highest purity; on the other hand, the trivial solution $N_p = 1$, which always yields a purity of 1, has also the lowest number of clusters. The trade-off is regulated by a purity threshold θ_{purity}, which has to be set by the user. The optimal N_p is chosen to be the largest among those corresponding to an average purity above θ_{purity}. Note that the solution $N_p = 1$ can be chosen, provided that purity at $N_p = 2$ is below θ_{purity}. To illustrate, Table B.3 shows how purity was computed for the example in Table B.2 with Eq. (B.9) five times for each candidate number of sub-phonemic clusters N_p, from 2 to $N_a = 5$, and then the optimal value $N_p = 3$ was chosen as the highest among the solutions above $\theta_{\text{purity}} = 0.75$.

N_p	Purity estimates					Average
2	0.95	0.95	1.0	1.0	0.9	0.96
3	0.95	0.95	0.95	0.95	0.95	**0.95**
4	0.58	0.50	0.50	0.50	0.50	0.52
5	0.50	-	0.50	0.50	0.50	-

Table B.3: Five computations of purity and their average value (columns) for different number of sub-phonemic clusters (rows). The optimal solution is $N_p = 3$, as its average is the one with the highest N_p among the solutions with average purity above $\theta_{\text{purity}} = 0.75$. The stochasticity of NMF allows for different solutions starting from the same input (rows). Missing values can occur whenever one or more clusters are empty, and such solutions are never chosen, as we consider them unstable.

B.5 Memory Management

Two measures were implemented so that the agents' memories are filled with an appropriate number of exemplars per word in the beginning and are kept at the same size throughout the simulation. In the case in which there are very few tokens from real speakers available for the initialisation of agents, the number of tokens can be increased by applying SMOTE (Chawla et al., 2002). This is a standard resampling technique that was adapted to the ABM architecture as explained in Appendix B.5.1. When the agents start interacting and accepting each others' exemplars, their memories will grow in size. In order to control the memory size, the model allows agent listeners to forget exemplars, i.e. when an agent listener has accepted and memorised a new exemplar, another exemplar of the same word class is removed from their memory. The only constraint on forgetting is that agents cannot forget word classes, i.e., removal is blocked if the deletion of an exemplar would lead to a decrease in the number of exemplars of the word class with respect to the number of exemplars with which the word class was initialised.

B.5.1 SMOTE

Synthetic Minority Over-sampling TEchnique (SMOTE, Chawla et al., 2002) is a non-parametric resampling algorithm used to mitigate the negative impact of imbalanced datasets on supervised classification. SMOTE has been applied to several classification problems (see Fernandez et al., 2018, for a survey), its performance and theoretical properties have been evaluated (Blagus & Lusa, 2013), and a number of modifications to the original algorithm have been proposed (Fernandez et al., 2018; Leevy et al., 2018). In its basic form, SMOTE generates extra artificial data for the under-represented class(es) in such a way as to preserve their original statistical properties. For example, if class A contains only 10 data points, while class B has 100, SMOTE can be applied to class A to top it up to the level of class B by generating 90 artificial data points. The generative process only makes use of the existing data points from the under-represented class. In other words, SMOTE does not require

the existence of a numerous class to generate extra data for a less numerous class. As a consequence, SMOTE can be applied more generally to increase the number of data points, e.g. when data are insufficient (though possibly balanced) for a given task.

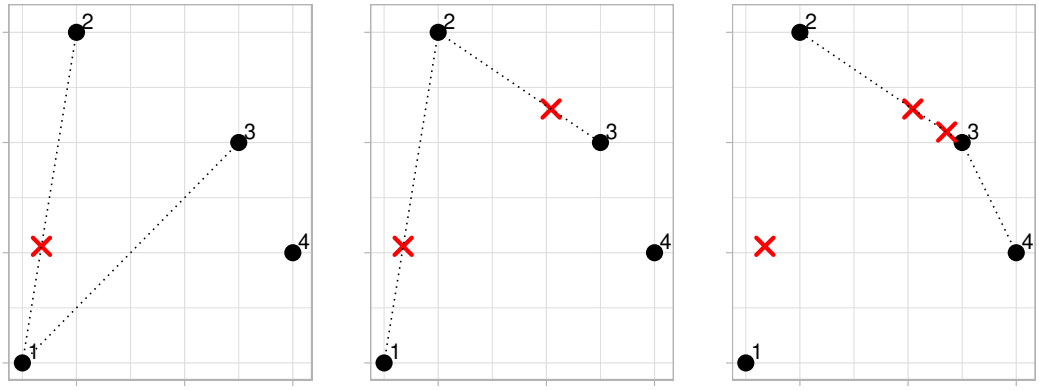


Figure B.3: Example of application of SMOTE to a set of four points (numbered). Each panel shows the generation of one new point (red ×'s), in succession from left to right. In each panel, a pivot point (point 1, 2, and 3 from left to right) is connected by segments to its $k = 2$ nearest neighbours. A new point is generated at a random position on one of those segments. The new points are not used in the generation process, i.e. they do not become pivot points nor are they considered as neighbours.

The data generation mechanism of SMOTE is illustrated in Figure B.3. Suppose there are n data points living in a d-dimensional space ($n = 4$ numbered points in a $d = 2$-dimensional space in Figure B.3) and n_S new points need to be created. Each new data point (red ×'s in Figure B.3) is generated by selecting a pivot point from the original n data points, connecting it to its k nearest neighbours ($k = 2$ in Figure B.3), selecting one of such connecting segments at random and generating a new data point along that segment at a random location. The process is repeated n_S times, each time changing the pivot point. Figure B.3 shows the generation of the first three points. Both k and n_S are parameters to be set by the user.

B.5.2 Application of SMOTE in the ABM

SMOTE has been employed in `soundChangeR` as a robustness measure for production. As described in section 3.2.3, the generation of a new acoustic token for speech production is implemented by first selecting a word class, then using all the exemplars of that word class available in the agent's memory to estimate a Gaussian distribution. The new token is produced by extracting one sample from that distribution. When the number of exemplars (data points) $N_e(w)$ of the target word class w is insufficient for a reliable estimation of a (multi-dimensional) Gaussian distribution, i.e. $N_e(w) < N_{\text{prod}}$, where N_{prod} is a user-defined threshold, $N_{\text{prod}} - N_e(w)$ extra exemplars for the target class are generated on the fly via SMOTE. After estimation, the extra exemplars are discarded. In case $N_e(w) < k+1$, i.e. if the number of available exemplars $N_e(w)$ is not even sufficient to identify k nearest neighbours, $k - N_e(w) + 1$ exemplars in the acoustic proximity of the $N_e(w)$ target exemplars from other word classes, but belonging to the same sub-phonemic class $p(w)$ as the target word, are added to the set of original data points used by SMOTE to generate new points.

The pure application of the above strategy was found to be insufficient as a protection against instability, especially at the beginning of ABM simulations. This is the case when very few tokens from real speakers are available for the initialisation of agents, so much so that not only production but also perception becomes unreliable, as the estimation of Gaussian mixtures necessary for the identification of acoustic clusters (cf. section 3.2.5 and Appendix B.4.1) is also negatively affected by data scarcity. To mitigate the problem, a strategy to populate agent memories based on production, thus ultimately based on SMOTE, was introduced. A threshold $N_{\text{ex/word}}$ defining the minimum number of exemplars per word per agent available at all times throughout a simulation was introduced. For all cases not meeting such requirement at the beginning of a simulation, i.e. if $N_e(w) < N_{\text{ex/word}}$ for some word w, $N_{\text{ex/word}} - N_e(w)$ extra exemplars are generated by the same mechanism used for production and stored in the agent's memory. To exemplify, suppose $N_e(w) = 3$ and $N_{\text{ex/word}} = 10$ for a given agent and word w at simulation start. Then 7 extra exemplars are produced, each one applying the ordinary production algorithm based

on the 3 original exemplars. If $N_e(w) < N_{\text{prod}}$, then SMOTE is applied each time, as explained above. Note that once a new exemplar is generated, it is not immediately made available as basis for the generation of the next ones. This is to avoid propagation of estimation errors. Only after all $N_{\text{ex/word}} - N_e(w)$ extra exemplars are generated, these are stored in memory and treated as original exemplars once the actual simulation (i.e. agent interactions) begins. To keep $N_e(w) \geq N_{\text{ex/word}}$ throughout the simulation, a constraint on the forgetting process was imposed such that no exemplar from a word class w can be forgotten (removed) if as a result the number of exemplars for w would go below the threshold $N_{\text{ex/word}}$ (cf. section 3.2.6). Note that $N_{\text{ex/word}}$ and N_{prod} are thresholds controlling different aspects, as the former is there to make sure that there is a minimum number of actual exemplars per word class in memory, which influences all aspects of the simulation, while the latter controls for the estimation basis for production only. In general $N_{\text{prod}} \geq N_{\text{ex/word}}$, for example if $N_{\text{ex/word}} = 10$ and $N_{\text{prod}} = 20$ it means that (i) $N_e(w) \geq 10$ for all w in all agents and (ii) SMOTE is applied in production whenever $10 \leq N_e(w) < 20$.

B.6 Production-Perception Feedback Loop

A schematic outline of the interaction between an agent speaker and listener is shown in Figure B.4, where both panels represent entities in a common two-dimensional acoustic space in the respective agents' memories. The agent speaker randomly chooses a word class, w_1, and builds a Gaussian model (black ellipse) over the acoustic representations of w_1's exemplars stored in its memory. The agent speaker then samples an acoustic value x from that model to build a new token (x, w_1) (w_1 in red, where x encodes its position in the Cartesian plane). The agent listener receives the token as is and looks up the phonological class with which w_1 is associated, in this case p_2. The agent listener memorises the token if it passes both the discriminability and typicality test.

The implementation of token production as well as the representation of sub-phonemic classes involve the use of density distributions. We opted for Gaussians and Gaussian mixture models because of their generality. In

particular, for production we opted for a unimodal Gaussian distribution, rather than a general GMM, because usually the number of tokens available for the estimation is quite reduced. Sub-phonemic classes are represented as a Mixture Discriminant Analysis model (MDA, Fraley & Raftery, 2002), a hierarchical model where each class is a GMM. For example, the listener panel in Figure B.4 represents an MDA model with three classes p_1–p_3, where p_1 is a two-component GMM, p_2 and p_3 are one-component GMMs, i.e. simple Gaussians.

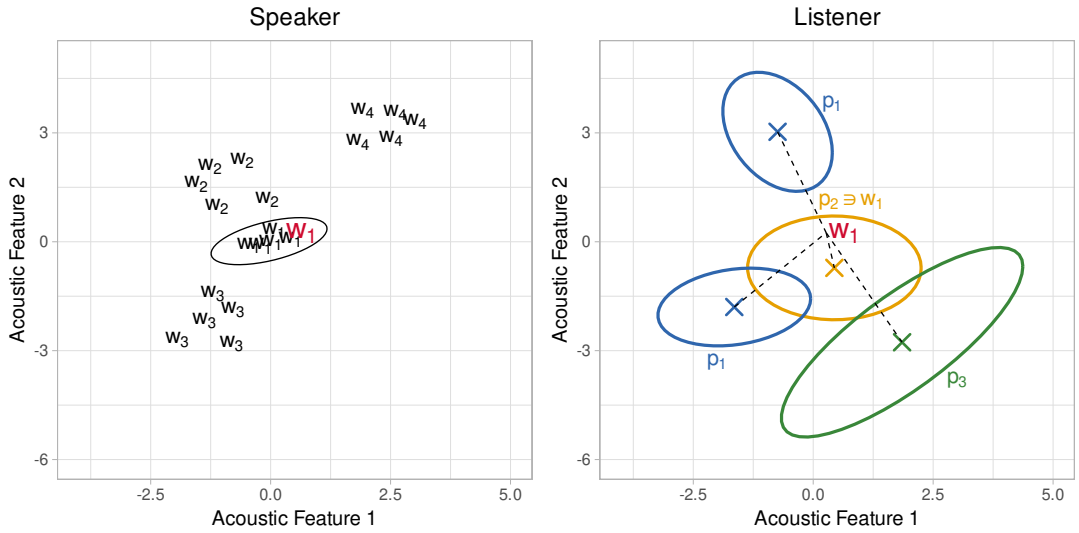


Figure B.4: Schematic representation of a token exchange between an agent speaker (left) and an agent listener (right). Ellipses represent mean-centred boundaries around Gaussian distributions enclosing 95% of their probability mass. The means are marked with a ×. The speaker produces a new token of a target sound of word type w_1 by estimating a Gaussian distribution based on all the stored exemplars of w_1 (black ellipse) and extracting a sample from it (in red). The listener receives the token as is and applies two memorisation tests based on its local phonological classes (coloured ellipses). The typicality test imposes a threshold on the maximum distance between the token and the phonological class it belongs to (here p_2); the discriminability test ensures that the token is closer to its phonological class than to the competing ones (here p_1 and p_3).

Typicality and discriminability tests involve some form of distance between a token and sub-phonemic classes. Typicality is operationalised by the Mahalanobis distances (Mahalanobis, 1936) between the token and all the GMM

components belonging to the relevant sub-phoneme. A threshold for the test was empirically set to a value corresponding to accepting tokens that lie within the mean-centered ellipsoid containing 95% of the Gaussian component probability mass. When the sub-phoneme consists of more than one component, the distances to all components are computed and the threshold is applied to the smallest distance. For example, the token received by the listener in Figure B.4 passes the typicality test because it lies inside the 95% probability mass ellipsoid of its class p_2, while it would fail if it belonged to class p_1 or p_3. The discriminability test is implemented by computing the maximum *a posteriori* probability (MAP) of the received token according to the MDA model. The test is passed if the sub-phonemic class the token belongs to according to its word label coincides with the MAP class; intuitively, the test is passed if the token is not acoustically ambiguous. For example, in Figure B.4 the received token (x, w_1), whose word label w_1 belongs to p_2, will pass the discriminability test if the posterior probability of x belonging to p_2 is higher than for the other two classes. More formally, a new exemplar x, which according to phonology, i.e. via word membership $p(w(e_i))$, belongs to phoneme y, is memorised only if Eq. (B.3) also yields y as class membership. Note that the MDA model is based on class memberships obtained via the chain $p(a(e_i))$, which in general does not yield the same result as the one that uses word membership, i.e. $p(w(e_i))$. The reason is that acoustic clusters contain impurities, i.e. exemplars that do not belong to the sub-phonemic class that is present as majority in a given acoustic cluster. We prefer to use this 'impure' association to build the classification model because this represents the acoustic counterpart of the sub-phonemic organisation of words, hence should be based on acoustic similarity rather than on pure word membership.

B.7 Implicit Assumptions, Proxies, and Simplifications

The agent-based model makes use of a number of implicit assumptions and simplifications, both in the way each agent is modelled as well as in the way

agents interact. Regarding the perception-production loop, first there is no explicit model of speech articulation. Any (co)articulation effect is indirectly modelled by the estimation of word-specific distributions, while sample extraction is a proxy for residual articulatory variation. Second, the transmission and perception of the acoustic properties of speech tokens between agents is assumed to be perfect. Besides not including any environmental acoustic noise, the ABM bypasses the acoustic domain entirely by implementing the acoustic perception of a token as a mere copy of its acoustic parameters, which implies that the identical acoustic parameterisation is assumed to be applied by all agents. The concept of token contains in itself the further assumption that all agents are parsing the continuous stream of speech in the same way insofar as the isolation of speech sounds is concerned. Finally, word recognition is assumed perfect, an assumption reflected by the error-free copy of the lexical information between interacting agents.

Agents' mental representations are also shaped by simplifying assumptions. The lexicon, identical for all agents, is actually a flat list of labels, as no orthographic or morphological information is used by the agents. More specifically, a "word" here is intended as the identification of a speech sound belonging to a certain word in a certain position, e.g. the vowel in *food*. There is no representation of a word as a sequence of speech sounds, which implies that there is no explicit or emerging pressure to maintain minimal pairs. The lexicon merely reflects the list of words of which agents have acoustic exemplars stored in memory, which in practice limits its size quite significantly. As a result, the emergence of sub-phonemic classes may be biased by the reduced variety of represented acoustic contexts.

There is no explicit notion of time. Implicitly, time advances at every interaction between two agents exchanging a token, which induces a partial ordering over interactions. In the current implementation, interactions occur only in pairs and only the agent listener can modify its state after an interaction by storing and/or forgetting an exemplar. Hence, if interaction 1 is a token exchange from agent A to B, interaction 2 from C to D, interaction 3 from C to E, and interaction 4 from B to F, then the only order constraint is that interaction 1 occurred before 4, because B as listener in 1 may have modified

its state before acting as speaker in 4, while all other interactions may have occurred in any order, e.g. $(1, 2, 3, 4)$ is indistinguishable from $(3, 1, 4, 2)$. The notion of time based on interactions is not calibrated, as we illustrate below. Each interaction involves the exchange of a single token, which may or may not be memorised by the listener. The impact of one new exemplar in memory on the future production and perception depends to a first approximation on the amount of exemplars already present in memory, i.e. the more the exemplars, the smaller the relative impact of a single new exemplar will be, as both perception and production are operationalised by some form of probability distribution estimation on the exemplars. However, the number of exemplars in an agent's memory is primarily determined by the available speech data, i.e. neither determined on the basis of any estimation on the amount of speech episodes humans may retain in memory, nor calibrated in such a way as to reproduce an experimentally quantified effect on speech production, e.g. based on imitation experiments. As a consequence, there is no way to determine an equivalence between, say, 1000 interactions and an amount of time in a real community of speakers. On top of this, interactions between agents are reduced to the exchange of a single token, while human conversations obviously involve exchange of a comparably larger amount of speech material between two (or more) speakers.

Agents' memories are initialised with speech material from actual speakers, typically one speaker initialising one agent, although the same speaker's material may be used to initialise more agents using resampling techniques. As a consequence, the amount of speakers available in the data set, which can be as small as 10 or 20, determines the size of the modelled agent population. A small population can introduce artefacts in that the relative importance of single individuals is disproportionately large. This may be a problem when the agent population models a large community of speakers, like in the cases presented here, while it may be adequate when modelling an actual isolated community (Harrington, Gubian et al., 2019). Groups of agents are obtained in two, possibly combined ways. First, groups of speakers of different accents or speaking styles are implicitly created by initialising agents with speech material from speakers of those distinct accents (cf. section 3.2.1), often using

an apparent time paradigm, whereby older and younger speakers of the same language are used as proxies for two different stages of sound change, the younger representing the more advanced one. Second, groups within a population can be predefined explicitly and thus be considered by the rules imposed on interactions, e.g. by allowing only inter- or only intra-group interactions. These groups may or may not coincide with the accent-based groups implicitly defined at initialisation.

C | Appendices to Chapter 4

C.1 Word List

/st/

vestuario – wardrobe
hasta – until
pistolín – small pistol
resto – rest
estaba – s/he/I was
estado – state
estanco – kiosk
pestaña – eyelash
destino – fate
estima – s/he respects
estío – summertime
pestiño – type of pastry
bestial – bestial
bestiando – (pseudoword)
destiempo – untimeliness
estuche – case
estufa – stove
estuve – I was
estuela – (pseudoword)

/t/

etapa – stage
retara – s/he/I may have challenged
etipa – (pseudoword)
retira – s/he takes away
returo – (pseudoword)
etupa – (pseudoword)
pata – paw
ata – s/he tied
repata – (pseudoword)

C.2 Median PC Score Values for Figure 4.3

Age	Region	Plosive	Median s_1	Median s_2	Median s_3	Median s_4
Older	East	/st/	0.103	−0.104	−0.062	−0.058
Older	East	/t/	−0.049	−0.015	−0.058	0.090
Older	West	/st/	0.065	−0.079	−0.024	−0.087
Older	West	/t/	−0.089	−0.042	−0.052	0.075
Younger	East	/st/	0.002	0.042	−0.058	−0.036
Younger	East	/t/	−0.074	0.078	−0.045	0.124
Younger	West	/st/	−0.154	0.034	0.051	−0.068
Younger	West	/t/	−0.156	−0.014	−0.016	0.132

Table C.1: Median PC score values used in Figure 4.3.

C.3 Median PC Score Values for Figure 4.4

Age	State	Plosive	Median s_1	Median s_2	Median s_3	Median s_4
Older	Baseline	/st/	0.088	−0.082	−0.038	−0.067
Older	Baseline	/t/	−0.064	−0.032	−0.059	0.072
Older	Post-run	/st/	0.032	−0.002	−0.034	−0.069
Older	Post-run	/t/	−0.090	−0.001	−0.042	0.131
Younger	Baseline	/st/	−0.073	0.045	−0.006	−0.049
Younger	Baseline	/t/	−0.109	0.022	−0.030	0.118
Younger	Post-run	/st/	0.028	−0.002	−0.034	−0.069
Younger	Post-run	/t/	−0.089	0.002	−0.044	0.131

Table C.2: Median PC score values used in Figure 4.4.

C.4 Estimated Marginal Means

Age	Region	Plosive	EMM	lower CL	higher CL
Older	East	/st/	0.122	0.029	0.215
Older	East	/t/	−0.019	−0.090	0.052
Older	West	/st/	0.113	0.020	0.206
Older	West	/t/	−0.079	−0.150	−0.008
Younger	East	/st/	0.057	−0.035	0.149
Younger	East	/t/	−0.063	−0.131	0.004
Younger	West	/st/	−0.151	−0.242	−0.059
Younger	West	/t/	−0.131	−0.198	−0.064

Table C.3: Estimated marginal means (EMM) with 95% confidence interval (lower and higher CL) computed by `emmeans` for the LMER with s_1 as dependent variable, for each combination of age group, region, and plosive.

Age	Region	Plosive	EMM	lower CL	higher CL
Older	East	/st/	−0.057	−0.143	0.029
Older	East	/t/	0.013	−0.079	0.105
Older	West	/st/	−0.063	−0.149	0.023
Older	West	/t/	−0.039	−0.131	0.053
Younger	East	/st/	0.061	−0.027	0.149
Younger	East	/t/	0.044	−0.052	0.140
Younger	West	/st/	0.055	−0.033	0.143
Younger	West	/t/	−0.008	−0.105	0.088

Table C.4: Estimated marginal means (EMM) with 95% confidence interval (lower and higher CL) computed by `emmeans` for the LMER with s_2 as dependent variable, for each combination of age group, region, and plosive.

Age	Region	EMM	lower CL	higher CL
Older	East	−0.038	−0.071	−0.005
Older	West	−0.004	−0.034	0.026
Younger	East	−0.005	−0.035	0.026
Younger	West	0.029	−0.001	0.060

Table C.5: Estimated marginal means (EMM) with 95% confidence interval (lower and higher CL) computed by `emmeans` for the LMER with s_3 as dependent variable, for each combination of age group and region.

Age	Region	Plosive	EMM	lower CL	higher CL
Older	East	/st/	−0.0582	−0.0817	−0.03475
Older	East	/t/	0.1050	0.0779	0.13217
Older	West	/st/	−0.0799	−0.1034	−0.05639
Older	West	/t/	0.0834	0.0562	0.11053
Younger	East	/st/	−0.0255	−0.0490	−0.00198
Younger	East	/t/	0.1378	0.1107	0.16493
Younger	West	/st/	−0.0471	−0.0706	−0.02363
Younger	West	/t/	0.1161	0.0890	0.14328

Table C.6: Estimated marginal means (EMM) with 95% confidence interval (lower and higher CL) computed by `emmeans` for the LMER with s_4 as dependent variable, for each combination of age group, region, and plosive.

Bibliography

Abramson, A. S. & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. In V. A. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 25–33). Academic Press. (Cit. on p. 126).

Abrego-Collier, C., Grove, J., Sonderegger, M. & Yu, A. C. L. (2011). Effects of speaker evaluation on phonetic convergence. *Proceedings of the 17th International Congress of the Phonetic Sciences*, 192–195 (cit. on p. 49).

Alderton, R. (2020a). Perceptions of T-glottalling among adolescents in South East England: A sign of 'chavviness', or a key to 'coolness'? *English Today*, *36*(3), 40–47 (cit. on p. 123).

Alderton, R. (2020b). Speaker Gender and Salience in Sociolinguistic Speech Perception: Goose-fronting in Standard Southern British English. *Journal of English Linguistics*, *48*(1), 72–96 (cit. on p. 93).

Alderton, R. (2022). T-tapping in Standard Southern British English: An 'elite' sociolinguistic variant? *Journal of Sociolinguistics*, *26*(2), 287–298 (cit. on p. 123).

Aleza Izquierdo, M. & Enguita Utrilla, J. M. (2002). *El español de América: Aproximación sincrónica*. Tirant lo Blanch. (Cit. on p. 16).

Alfonso, P. J. & Baer, T. (1982). Dynamics of Vowel Articulation. *Language and Speech*, *25*(2), 151–173 (cit. on p. 18).

Allen, J. S., Miller, J. L. & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552 (cit. on p. 1).

Alvar, M. (1955). Las hablas meridionales de España y su interés para la lingüística comparada. *Revista de Filología Española*, *39*(1/4), 284–313 (cit. on p. 15).

Asano, Y. & Gubian, M. (2018). ''Excuse meeee!!": (Mis)coordination of lexical and paralinguistic prosody in L2 hyperarticulation. *Speech Communication*, *99*, 183–200 (cit. on p. 158).

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177–189 (cit. on p. 49).

Babel, M., McGuire, G., Walters, S. & Nicholls, A. (2014). Novelty and social preference in phonetic accommodation. *Laboratory Phonology*, *5*(1), 123–150 (cit. on p. 2).

Bailey, G. (2017). Field Interviews in Dialectology. In C. Boberg, J. Nerbonne & D. Watt (Eds.), *The Handbook of Dialectology* (pp. 284–299). Wiley. (Cit. on p. 23).

Bailey, G., Wikle, T., Tillery, J. & Sand, L. (1991). The apparent time construct. *Language Variation and Change*, *3*, 241–264 (cit. on pp. viii, xiv, 78, 88, 108).

Baker, A., Archangeli, D. & Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change*, *23*(3), 347–374 (cit. on pp. 4, 47, 123).

Bang, H.-Y., Sonderegger, M., Kang, Y., Clayards, M. & Yoon, T.-J. (2018). The emergence, progress, and impact of sound change in progress in Seoul Korean: Implications for mechanisms of tonogenesis. *Journal of Phonetics*, *66*, 120–144 (cit. on p. 125).

Bankes, S. C. (2002). Agent-based modeling: A revolution? *Proceedings of the National Academy of Sciences*, *99*, 7199–7200 (cit. on p. 46).

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48 (cit. on p. 112).

Beck, J. M. (2010). Organic Variation of the Vocal Apparatus. In W. J. Hardcastle, J. Laver & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (pp. 153–201). Blackwell Publishing Ltd. (Cit. on p. 1).

Beckman, M. E., De Jong, K., Jun, S.-A. & Lee, S.-H. (1992). The Interaction of Coarticulation and Prosody in Sound Change. *Language and Speech*, *35*(1-2), 45–58 (cit. on p. 4).

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D. & Schoenemann, T. (2009). Language Is a Complex Adaptive System: Position Paper. *Language Learning*, *59*, 1–26 (cit. on pp. 46, 64).

Beddor, P. S. (2009). A Coarticulatory Path to Sound Change. *Language*, *85*(4), 785–821 (cit. on pp. ix, xv, 4, 17–19, 41, 42, 92, 93, 125, 139).

Beddor, P. S. (2012). Perception Grammars and Sound Change. In M.-J. Solé & D. Recasens (Eds.), *The Initiation of Sound Change: Perception, Production, and Social Factors* (pp. 37–56). John Benjamins. (Cit. on pp. 18, 19, 41).

Beddor, P. S., Brasher, A. & Narayan, C. (2007). Applying Perceptual Methods to the Study of Phonetic Variation and Sound Change. In M.-J. Solé, P. S. Beddor & M. Ohala (Eds.), *Experimental Approaches to Phonology* (pp. 127–143). Oxford University Press. (Cit. on p. 19).

Beddor, P. S., Coetzee, A. W., Styler, W., McGowan, K. B. & Boland, J. E. (2018). The time course of individuals' perception of coarticulatory information is linked to their production: Implications for sound change. *Language*, *94*(4), 931–968 (cit. on pp. 18, 41, 78).

Bermúdez-Otero, R. (2015). Amphichronic Explanation and the Life Cycle of Phonological Processes. In P. Honeybone & J. C. Salmons (Eds.), *The Oxford Handbook of Historical Phonology* (pp. 374–399). Oxford University Press. (Cit. on p. 19).

Bermúdez-Otero, R. (2020). The initiation and incrementation of sound change: Community-oriented momentum-sensitive learning. *Glossa*, *5*(1), 1–32 (cit. on pp. 4, 47).

Bermúdez-Otero, R. & Trousdale, G. (2012). Cycles and continua: On unidirectionality and gradualness in language change. In T. Nevalainen & E. Closs Traugott (Eds.), *The Oxford Handbook of the History of English* (pp. 691–720). Oxford University Press. (Cit. on pp. 19, 139).

Berry, B. J. L., Kiel, L. D. & Elliott, E. (2002). Adaptive agents, intelligence, and emergent human organization: Capturing complexity through agent-based modeling. *Proceedings of the National Academy of Sciences*, *99*, 7187–7188 (cit. on p. 46).

Best, C. T., Morrongiello, B. & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, *29*(3), 191–211 (cit. on p. 18).

Bladon, A. (1986). Phonetics for hearers. In G. McGregor (Ed.), *Language for hearers* (pp. 1–24). Pergamon. (Cit. on pp. 6, 16, 93).

Blagus, R. & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(1), 106 (cit. on p. 202).

Blevins, J. & Wedel, A. (2009). Inhibited sound change: An evolutionary approach to lexical competition. *Diachronica*, *26*(2), 143–183 (cit. on pp. 4, 77, 79, 80, 82, 83, 94).

Boersma, P., Escudero, P. & Hayes, R. (2003). Learning Abstract Phonological from Auditory Phonetic Categories: An Integrated Model for the Acquisition of Language-Specific Sound Categories. *Proceedings of the 15th International Congress of Phonetic Sciences*, 1013–1016 (cit. on p. 18).

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*, 7280–7287 (cit. on p. 46).

Browman, C. P. & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219–252 (cit. on pp. 20, 92, 96, 121).

Browman, C. P. & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201–251 (cit. on pp. 8, 92, 121).

Browman, C. P. & Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica*, *49*(3-4), 155–180 (cit. on pp. 8, 20, 96).

Butterworth, S. (1930). On the Theory of Filter Amplifiers. *Wireless Engineer*, *7*(6), 536–541 (cit. on p. 23).

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*(3), 261–290 (cit. on p. 85).

Bybee, J. (2015). Articulatory Processing and Frequency of Use in Sound Change. In P. Honeybone & J. Salmons (Eds.), *The Oxford Handbook of Historical Phonology* (pp. 467–484). Oxford University Press. (Cit. on pp. 4, 85).

Campeanu, S., Craik, F. I. M., Backer, K. C. & Alain, C. (2014). Voice reinstatement modulates neural indices of continuous word recognition. *Neuropsychologia*, *62*, 233–244 (cit. on p. 2).

Canfield, D. L. (1981). *Spanish pronunciation in the Americas*. University of Chicago Press. (Cit. on pp. vii, xiii, 14).

Carignan, C., Coretta, S., Frahm, J., Harrington, J., Hoole, P., Joseph, A., Kunay, E. & Voit, D. (2021). Planting the seed for sound change: Evidence from

real-time MRI of velum kinematics in German. *Language*, *97*(2), 333–364 (cit. on p. 93).

Carignan, C., Hoole, P., Kunay, E., Joseph, A., Voit, D., Frahm, J. & Harrington, J. (2019). The phonetic basis of phonological vowel nasality: Evidence from real-time MRI velum movement in German. *Proceedings of the 19th International Congress of Phonetic Sciences* (cit. on pp. 92, 125).

Castiglione, F., Deb, D., Srivastava, A. P., Liò, P. & Liso, A. (2021). From Infection to Immunity: Understanding the Response to SARS-CoV2 Through In-Silico Modeling. *Frontiers in Immunology*, *12*, 646972 (cit. on p. 5).

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357 (cit. on pp. 61, 202).

Cheng, C.-H., Hsu, W.-Y., Shih, Y.-H., Lin, H.-C., Liao, K.-K., Wu, Z.-A. & Lin, Y.-Y. (2010). Differential cerebral reactivity to shortest and longer tones: Neuromagnetic and behavioral evidence. *Hearing Research*, *268*(1-2), 260–270 (cit. on p. 123).

Church, B. A. & Schacter, D. L. (1994). Perceptual Specificity of Auditory Priming: Implicit Memory for Voice Intonation and Fundamental Frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 521–533 (cit. on p. 2).

Clark, L. & Trousdale, G. (2009). Exploring the role of token frequency in phonological change: Evidence from TH-Fronting in east-central Scotland. *English Language and Linguistics*, *13*(1), 33–55 (cit. on p. 85).

Clarke-Davidson, C. M., Luce, P. A. & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception & Psychophysics*, *70*(4), 604–618 (cit. on pp. 2, 50).

Clayards, M. (2018). Differences in cue weights for speech perception are correlated for individuals within and across contrasts. *The Journal of the Acoustical Society of America*, *144*(3), EL172–EL177 (cit. on pp. 18, 126).

Clopper, C. G. (2014). Sound change in the individual: Effects of exposure on cross-dialect speech processing. *Laboratory Phonology*, *5*(1), 69–90 (cit. on p. 78).

Clopper, C. G., Pierrehumbert, J. B. & Tamati, T. N. (2010). Lexical neighborhoods and phonological confusability in cross-dialect word recognition in noise. *Laboratory Phonology*, *1*(1), 65–92 (cit. on p. 86).

Clopper, C. G. & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, *32*(1), 111–140 (cit. on p. 123).

Clopper, C. G. & Pisoni, D. B. (2005). Perception of Dialect Variation. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 313–337). Blackwell. (Cit. on p. 123).

Coetzee, A. W., Beddor, P. S., Shedden, K., Styler, W. & Wissing, D. (2018). Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis. *Journal of Phonetics*, *66*, 185–216 (cit. on pp. 92, 125).

Cohn, A. C. (2007). Phonetics in Phonology and Phonology in Phonetics. *Working Papers of the Cornell Phonetics Laboratory*, *16*, 1–31 (cit. on p. 127).

Coleman, J. (2002). Phonetic Representations in the Mental Lexicon. In J. Durand & B. Laks (Eds.), *Phonetics, Phonology, and Cognition* (pp. 69–130). Oxford University Press. (Cit. on p. 82).

Connine, C. M. (2004). It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. *Psychonomic Bulletin & Review*, *11*(6), 1084–1089 (cit. on p. 86).

Connine, C. M. & Darnieder, L. M. (2009). Perceptual learning of co-articulation in speech. *Journal of Memory and Language*, *61*(3), 412–422 (cit. on p. 2).

Connine, C. M., Titone, D. & Wang, J. (1993). Auditory Word Recognition: Extrinsic and Intrinsic Effects of Word Frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 81–94 (cit. on p. 86).

Creel, S. C. & Bregman, M. R. (2011). How Talker Identity Relates to Language Processing: Talkers and Language Processing. *Language and Linguistics Compass*, *5*(5), 190–204 (cit. on p. 124).

Cronenberg, J., Gubian, M., Harrington, J. & Ruch, H. (2020). A dynamic model of the change from pre- to post-aspiration in Andalusian Spanish. *Journal of Phonetics*, *83*, 101016 (cit. on pp. viii, xiv, 13, 92, 93, 96–98, 100, 102, 108, 109, 121, 130, 136, 139, 143).

Cronenberg, J., Gubian, M., Harrington, J. & Schiel, F. (2022). *soundChangeR: An Agent-Based Model of Sound Change* [R package version 1.0.0]. (Cit. on p. 94).

Cutler, A., Eisner, F., McQueen, J. M. & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, M. d'Imperio & N. Vallée (Eds.), *Laboratory Phonology 10* (pp. 91–111). Mouton de Gruyter. (Cit. on p. 64).

Dahan, D., Magnuson, J. S. & Tanenhaus, M. K. (2001). Time Course of Frequency Effects in Spoken-Word Recognition: Evidence from Eye Movements. *Cognitive Psychology*, *42*(4), 317–367 (cit. on p. 86).

Daikhin, L. & Ahissar, M. (2012). Responses to deviants are modulated by subthreshold variability of the standard. *Psychophysiology*, *49*(1), 31–42 (cit. on p. 123).

Davidson, L. (2006). Schwa Elision in Fast Speech: Segmental Deletion or Gestural Overlap? *Phonetica*, *63*(2-3), 79–112 (cit. on p. 17).

Dediu, D. & Moisik, S. R. (2019). Pushes and pulls from below: Anatomical variation, articulation and sound change. *Glossa*, *4*(1), 1–33 (cit. on p. 51).

Del Saz, M. (2019). From postaspiration to affrication: New phonetic contexts in Western Andalusian Spanish. *Proceedings of the 19th International Congress of Phonetic Sciences*, 760–764 (cit. on pp. 139, 141).

Delvaux, V. & Soquet, A. (2007). The Influence of Ambient Speech on Adult Speech Productions through Unintentional Imitation. *Phonetica*, *64*(2-3), 145–173 (cit. on p. 49).

Ding, C., He, X. & Simon, H. D. (2005). On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. *Proceesings of the 5th SIAM International Conference on Data Mining*, 606–610 (cit. on p. 193).

Divjak, D. (2019). *Frequency in Language: Memory, Attention and Learning*. Cambridge University Press. (Cit. on p. 85).

Draxler, C. & Jänsch, K. (2004). SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, 559–562 (cit. on p. 97).

Eckert, P. (1988). Adolescent social structure and the spread of linguistic change. *Language in Society*, *17*(2), 183–207 (cit. on p. 47).

Eckert, P. (1989). The whole woman: Sex and gender differences in variation. *Language Variation and Change*, *1*, 245–267 (cit. on p. 2).

Egurtzegi, A., Cronenberg, J., Gubian, M., Harrington, J. & Ruch, H. (2022). From post-lexicality to phonologization: Andalusian /s/-aspiration at the word boundary. *Proceedings of the 18th Conference on Laboratory Phonology*, 1–2 (cit. on p. 139).

Eisner, F. & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*(4), 1950–1953 (cit. on pp. 2, 50).

Eisner, F. & Mcqueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238 (cit. on p. 2).

Ettlinger, M. (2007). An exemplar-based model of chain shifts. *Proceedings of the 16th International Congress of Phonetic Sciences*, 685–688 (cit. on pp. 39, 77, 82, 83, 85, 128).

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L. & Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua*, *120*(8), 2061–2079 (cit. on pp. 79, 94).

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton de Gruyter. (Cit. on p. 1).

Fant, G. (1973). *Speech Sounds and Features*. MIT Press. (Cit. on p. 16).

Fenn, K. M., Nusbaum, H. C. & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, *425*(6958), 614–616 (cit. on p. 2).

Fernandez, A., Garcia, S., Herrera, F. & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905 (cit. on p. 202).

Fitch, H. L., Halwes, T., Erickson, D. & Liberman, A. M. (1979). The perceptual equivalence of trading-relation cues. *The Journal of the Acoustical Society of America*, *65*(S1), 80 (cit. on p. 93).

Flege, J. E., Birdsong, D., Bialystok, E., Mack, M., Sung, H. & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics*, *34*, 153–175 (cit. on p. 87).

Forster, K. I. & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*(6), 627–635 (cit. on p. 86).

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics*, *36*(4), 359–368 (cit. on pp. viii, xiv, 18, 39).

Fowler, C. A. (2003). Speech production and perception. In I. B. Weiner, A. F. Healy & R. W. Proctor (Eds.), *Handbook of Psychology. Volume 4: Experimental Psychology.* (pp. 237–266). Wiley. (Cit. on p. 43).

Fowler, C. A. (2005). Parsing coarticulated speech in perception: Effects of coarticulation resistance. *Journal of Phonetics*, *33*(2), 199–213 (cit. on pp. 18, 24).

Fowler, C. A. & Brown, J. M. (2000). Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Perception and Psychophysics*, *62*(1), 21–32 (cit. on p. 24).

Fowler, C. A. & Saltzman, E. (1993). Coordination and Coarticulation in Speech Production. *Language and Speech*, *36*(2-3), 171–195 (cit. on p. 38).

Fowler, C. A. & Smith, M. R. (1986). Speech Perception as Vector Analysis: An Approach to the Problem of Invariance and Segmentation. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and Variability in Speech Processes* (pp. 123–136). Lawrence Erlbaum Associates. (Cit. on pp. 8, 20, 38, 39).

Fraley, C. & Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, *97*(458), 611–631 (cit. on pp. 191, 206).

Francis, A. L., Baldwin, K. & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics*, *62*(8), 1668–1680 (cit. on pp. 18, 93, 126).

Francis, A. L., Kaganovich, N. & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *The Journal of the Acoustical Society of America*, *124*(2), 1234–1251 (cit. on p. 126).

Francis, A. L. & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 349–366 (cit. on pp. 18, 126).

Fridland, V. (2008). Patterns of /uw/, /U/, and /ow/ fronting in Reno, Nevada. *American Speech*, *83*(4), 432–454 (cit. on p. 93).

Fruehwald, J. (2017). The Role of Phonology in Phonetic Change. *Annual Review of Linguistics*, *3*(1), 25–42 (cit. on p. 139).

Garellek, M., Simpson, A., Roettger, T. B., Recasens, D., Niebuhr, O., Mooshammer, C., Michaud, A., Lee, W.-S., Kirby, J., Gordon, M. & Yu, K. M. (2020). Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls. *Journal of Speech Sciences*, *9*, 3–16 (cit. on p. 89).

Garnier, M., Lamalle, L. & Sato, M. (2013). Neural correlates of phonetic convergence and speech imitation. *Frontiers in Psychology*, *4*, 1–15 (cit. on p. 49).

Garrett, A. & Johnson, K. (2013). Phonetic bias in sound change. In A. C. L. Yu (Ed.), *Origins of Sound Change* (pp. 51–97). Oxford University Press. (Cit. on pp. 2, 48, 123, 125).

Garrido, M. I., Sahani, M. & Dolan, R. J. (2013). Outlier Responses Reflect Sensitivity to Statistical Structure in the Human Brain (O. Sporns, Ed.). *PLoS Computational Biology*, *9*(3), e1002999 (cit. on p. 122).

Gaujoux, R. & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, *11*(1), 367 (cit. on p. 193).

Gerfen, C. (2002). Andalusian codas. *Probus*, *14*(2), 247–277 (cit. on p. 15).

German, J. S., Carlson, K. & Pierrehumbert, J. B. (2013). Reassignment of consonant allophones in rapid dialect acquisition. *Journal of Phonetics*, *41*(3-4), 228–248 (cit. on p. 51).

Gilbert, M. (2022). *An experimental and formal investigation of Sevillian Spanish metathesis* (Doctoral dissertation). New York University. New York. (Cit. on pp. 92, 108).

Gilbert, M. (2023a). Testing for underlying representations: Segments and clusters in Sevillian Spanish. *Natural Language & Linguistic Theory*, 1–39 (cit. on p. 92).

Gilbert, M. (2023b). Testing the perceptual basis of laryngeal metathesis and rarity of preaspirated stops. *Proceedings of the 20th International Congress of Phonetic Sciences*, 117–121 (cit. on pp. 6, 93).

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics*, *15*(2), 87–105 (cit. on p. 49).

Goldinger, S. D. (1996). Words and Voices: Episodic Traces in Spoken Word Identification and Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183 (cit. on pp. ix, xv, 2, 50).

Goldinger, S. D. (1998). Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review*, *105*(2), 251–279 (cit. on pp. 2, 50).

Goldinger, S. D. & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review*, *11*(4), 716–722 (cit. on pp. ix, xv, 2, 96).

Goldstein, L. & Browman, C. P. (1986). Representation of voicing contrasts using articulatory gestures. *Journal of Phonetics*, *14*, 339–342 (cit. on pp. 8, 98, 108).

Goldstone, R. L. (1998). Perceptual Learning. *Annual Review of Psychology*, *49*, 585–612 (cit. on pp. 2, 126).

Gomez, J., Prieto, J., Leon, E. & Rodríguez, A. (2021). INFEKTA – An agent-based model for transmission of infectious diseases: The COVID-19 case in Bogotá, Colombia (R. L. Smith, Ed.). *PLoS ONE*, *16*(2), e0245787 (cit. on p. 5).

Gonzalez, S. & Brookes, M. (2014). PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(2), 518–530 (cit. on pp. viii, xiv, 23).

Gordon, M. (2013). Investigating Chain Shifts and Mergers. In J. K. Chambers & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (2nd ed., pp. 203–219). Wiley-Blackwell. (Cit. on p. 79).

Greca, P., Gubian, M. & Harrington, J. (2022). The relationship between the coarticulatory source and effect in sound change: Evidence from Italo-Romance metaphony in the Lausberg area. *Laboratory Phonology* (cit. on pp. 93, 125, 127).

Gubian, M., Boves, L. & Cangemi, F. (2011). Joint analysis of f0 and speech rate with Functional Data Analysis. *Proceedings of the 36th International Conference on Acoustics, Speech, and Signal Processing*, 4972–4975 (cit. on p. 158).

Gubian, M., Cronenberg, J. & Harrington, J. (2023). Phonetic and Phonological Sound Changes in an Agent-Based Model. *Speech Communication*, *147*, 93–115 (cit. on pp. xi, xvi, 45, 58, 59, 68, 73, 81, 82, 112, 113, 135, 165).

Gubian, M., Harrington, J., Stevens, M., Schiel, F. & Warren, P. (2019). Tracking the New Zealand English NEAR/SQUARE merger using functional principal components analysis. *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, 296–300 (cit. on p. 21).

Gubian, M., Torreira, F. & Boves, L. (2015). Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, *49*, 16–40 (cit. on pp. 21, 24, 27, 99).

Hackl, J. & Dubernet, T. (2019). Epidemic Spreading in Urban Areas Using Agent-Based Transportation Models. *Future Internet*, *11*(4), 92–108 (cit. on p. 5).

Hagège, C. & Haudricourt, A.-G. (1978). *La phonologie panchronique: Comment les sons changent dans les langues*. Presses universitaires de France. (Cit. on pp. 18, 125).

Haggard, M., Summerfield, Q. & Roberts, M. (1981). Psychoacoustical and cultural determinants of phoneme boundaries: Evidence from trading F0 cues in the voiced-voiceless distinction. *Journal of Phonetics*, *9*(1), 49–62 (cit. on pp. 18, 93).

Hall-Lew, L., Coppock, E. & Starr, R. L. (2010). Indexing political persuasion: Variation in the Iraq vowels. *American Speech*, *85*(1), 91–102 (cit. on p. 2).

Harmon, Z., Idemaru, K. & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition*, *189*, 76–88 (cit. on pp. 18, 125, 126).

Harrington, J. (2014). Variability and change in spoken language communication. In E. Glaser, A. Kolmer, M. Meyer & E. Stark (Eds.), *Sprache(n) verstehen* (pp. 33–57). Hochschulverlag ETH Zürich. (Cit. on pp. 1, 78).

Harrington, J. & Cassidy, S. (1999). *Techniques in Speech Acoustics*. Springer Netherlands. (Cit. on p. 23).

Harrington, J., Gubian, M., Stevens, M. & Schiel, F. (2019). Phonetic change in an Antarctic winter. *The Journal of the Acoustical Society of America*, *146*(5), 3327–3332 (cit. on pp. 89, 94, 135, 209).

Harrington, J., Kleber, F. & Reubold, U. (2008). Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *The Journal of the Acoustical Society of America*, *123*(5), 2825–2835 (cit. on pp. 58, 83, 93, 198).

Harrington, J., Kleber, F. & Reubold, U. (2012). The production and perception of coarticulation in two types of sound changes in progress. In S. Fuchs, M. Weirich, D. Pape & P. Perrier (Eds.), *Speech Planning and Dynamics* (pp. 33–55). Peter Lang. (Cit. on p. 4).

Harrington, J., Kleber, F., Reubold, U., Schiel, F. & Stevens, M. (2018). Linking Cognitive and Social Aspects of Sound Change Using Agent-Based Modeling. *Topics in Cognitive Science*, 1–22 (cit. on pp. ix, xv, 16, 47, 49, 51, 64, 80, 88, 121, 123, 127, 131).

Harrington, J., Kleber, F., Reubold, U., Schiel, F. & Stevens, M. (2019). The phonetic basis of the origin and spread of sound change. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge Handbook of Phonetics* (pp. 401–426). Routledge. (Cit. on p. 4).

Harrington, J. & Reubold, U. (2021). Accent Reversion in Older Adults: Evidence from the Queen's Christmas Broadcasts. In K. V. Beaman & I. Buchstaller (Eds.), *Language Variation and Language Change Across the Lifespan: Theoretical and Empirical Perspectives from Panel Studies* (1st ed., pp. 119–137). Routledge. (Cit. on p. 85).

Harrington, J. & Schiel, F. (2017). /u/-fronting and agent-based modeling: The relationship between the origin and spread of sound change. *Language*, *93*(2), 414–445 (cit. on pp. 47, 52, 57, 58, 61, 84, 94, 121, 123, 135, 198, 200).

Harrington, J. & Stevens, M. (2014). Cognitive processing as a bridge between phonetic and social models of sound change. *Laboratory Phonology*, *5*(1), 1–8 (cit. on p. 4).

Hassani-Mahmooei, B. & Parris, B. W. (2012). Climate change and internal migration patterns in Bangladesh: An agent-based model. *Environment and Development Economics*, *17*(6), 763–780 (cit. on p. 5).

Hay, J. B. & Foulkes, P. (2016). The evolution of medial /t/ over real and remembered time. *Language*, *92*(2), 298–330 (cit. on p. 85).

Hay, J. B., Jannedy, S. & Mendoza-Denton, N. (1999). Oprah and /ay/: Lexical Frequency, Referee Design and Style. *Proceedings of the 14th International Congress of Phonetic Sciences*, 1389–1392 (cit. on p. 2).

Hay, J. B., Nolan, A. & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review*, *23*, 351–379 (cit. on p. 124).

Hay, J. B., Pierrehumbert, J. B., Walker, A. J. & LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition*, *139*, 83–91 (cit. on pp. 79, 85, 86).

Henriksen, N. & Harper, S. K. (2016). Investigating lenition patterns in south-central Peninsular Spanish /sp, st, sk/ clusters. *Journal of the International Phonetic Association*, *46*(3), 287–310 (cit. on p. 141).

Herrero de Haro, A. (2017). The phonetics and phonology of Eastern Andalusian Spanish: A review of literature from 1881 to 2016. *Revista de Lenguaje y Cultura*, *22*(2), 313–357 (cit. on p. 15).

Holt, L. L. & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071 (cit. on pp. 18, 93).

Hombert, J.-M., Ohala, J. J. & Ewan, W. G. (1979). Phonetic Explanations for the Development of Tones. *Language*, *55*(1), 37–58 (cit. on p. 18).

Hooper, J. (1976). Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie (Ed.), *Current Progress in Historical Linguistics* (pp. 96–105). North Holland. (Cit. on p. 85).

Hualde, J. I. & Chitoran, I. (2016). Surface sound and underlying structure: The phonetics-phonology interface. In S. Fischer & C. Gabriel (Eds.), *Manual of Grammatical Interfaces in Romance* (pp. 23–40). De Gruyter. (Cit. on p. 15).

Hualde, J. I. & Sanders, B. P. (1995). A New Hypothesis on the Origin of the Eastern Andalusian Vowel System. *Proceedings of the 21st Annual Meeting of the Berkeley Linguistics Society*, 426–437 (cit. on p. 15).

Hyman, L. M. (1976). Phonologization. In A. Juilland (Ed.), *Linguistic studies presented to Joseph H. Greenberg* (pp. 407–418). Anma Libri. (Cit. on p. 125).

Hyman, L. M. (2013). Enlarging the scope of phonologization. In A. C. L. Yu (Ed.), *Origins of Sound Change* (pp. 3–28). Oxford University Press. (Cit. on pp. 19, 127).

Idemaru, K. & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–1956 (cit. on pp. 18, 123).

Idemaru, K., Holt, L. L. & Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America*, *132*(6), 3950–3964 (cit. on pp. 18, 126).

Jacewicz, E., Arzbecker, L. J. & Fox, R. A. (2021). Perception of indexical cues in speech by children and adults with and without dyslexia: Regional dialect and gender identification. *Dyslexia*, *28*(1), 60–78 (cit. on p. 123).

Jescheniak, J. D. & Levelt, W. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824–843 (cit. on p. 86).

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 145–165). Academic Press. (Cit. on pp. 2, 39, 124).

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *34*, 485–499 (cit. on pp. 2, 39, 124).

Johnson, K., Ladefoged, P. & Lindau, M. (1993). Individual differences in vowel production. *The Journal of the Acoustical Society of America*, *94*(2), 701–714 (cit. on p. 2).

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R2 GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946 (cit. on p. 157).

Jones, Z. & Clopper, C. G. (2019). Subphonemic Variation and Lexical Processing: Social and Stylistic Factors. *Phonetica*, *76*(2-3), 163–178 (cit. on p. 51).

Kang, Y. (2014). Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: A corpus study. *Journal of Phonetics*, *45*, 76–90 (cit. on pp. 94, 125).

Kang, Y. & Han, S. (2013). Tonogenesis in early Contemporary Seoul Korean: A longitudinal case study. *Lingua*, *134*, 62–74 (cit. on pp. 92, 94, 125).

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J. & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(9), 1594–1611 (cit. on p. 126).

Kappes, J., Baumgärtner, A., Peschke, C. & Ziegler, W. (2009). Unintended imitation in nonword repetition. *Brain and Language*, *111*(3), 140–151 (cit. on p. 49).

Kawasaki, M., Yamada, Y., Ushiku, Y., Miyauchi, E. & Yamaguchi, Y. (2013). Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction. *Scientific Reports*, *3*(1), 1692 (cit. on p. 50).

Kelso, J. A. S. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *246*(6), 87–95 (cit. on p. 93).

Kerswill, P. & Trudgill, P. (2005). The birth of new dialects. In P. Auer, F. Hinskens & P. Kerswill (Eds.), *Dialect Change: Convergence and Divergence in European Languages* (pp. 169–220). Cambridge University Press. (Cit. on pp. 4, 87).

Keskinocak, P., Oruc, B. E., Baxter, A., Asplund, J. & Serban, N. (2020). The impact of social distancing on COVID19 spread: State of Georgia case study (M. Adrish, Ed.). *PLoS ONE*, *15*(10), e0239798 (cit. on p. 5).

Kim, D., Clayards, M. & Kong, E. J. (2017). Individual differences in perceptual adaptation to phonetic categories: Categorization gradiency and cognitive abilities. *The Journal of the Acoustical Society of America*, *142*(4), 2704 (cit. on p. 125).

Kim, J. & Park, H. (2008). *Sparse Nonnegative Matrix Factorization for Clustering* (tech. rep. GT-CSE-08-01). Georgia Institute of Technology. Atlanta. (Cit. on p. 193).

Kingston, J. (2011). Tonogenesis. In M. van Oostendorp, C. J. Ewen, E. V. Hume & K. Rice (Eds.), *The Blackwell Companion to Phonology* (pp. 2304–2333). Wiley-Blackwell. (Cit. on p. 92).

Kingston, J., Diehl, R. L., Kirk, C. J. & Castleman, W. A. (2008). On the internal perceptual structure of distinctive features: The [voice] contrast. *Journal of Phonetics*, *36*(1), 28–54 (cit. on p. 18).

Kiparsky, P. (2015). Phonologization. In P. Honeybone & J. Salmons (Eds.), *The Oxford Handbook of Historical Phonology* (pp. 563–582). Oxford University Press. (Cit. on pp. 19, 93, 139).

Kiparsky, P. (2018). Formal and empirical issues in phonological typology. In L. M. Hyman & F. Plank (Eds.), *Phonological Typology* (pp. 54–106). De Gruyter Mouton. (Cit. on p. 82).

Kirby, J. (2013). The role of probabilistic enhancement in phonologization. In A. C. L. Yu (Ed.), *Origins of Sound Change* (pp. 228–246). Oxford University Press. (Cit. on pp. 4, 78, 79, 82, 83, 85, 92, 94, 125–128).

Kirby, J. (2014a). Incipient tonogenesis in Phnom Penh Khmer: Acoustic and perceptual studies. *Journal of Phonetics*, *43*, 69–85 (cit. on pp. 92, 95).

Kirby, J. (2014b). Incipient tonogenesis in Phnom Penh Khmer: Computational studies. *Laboratory Phonology*, *5*(1), 195–230 (cit. on pp. 18, 77–79, 86, 94, 95, 125, 127, 128).

Kirby, J. & Sonderegger, M. (2013). A model of population dynamics applied to phonetic change. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 776–781 (cit. on pp. 77, 86, 94).

Kirby, J. & Sonderegger, M. (2015). Bias and population structure in the actuation of sound change. *arXiv*, 1–30 (cit. on pp. 4, 48, 94).

Kirchner, R., Moore, R. K. & Chen, T.-Y. (2010). Computing phonological generalization over real speech exemplars. *Journal of Phonetics*, *38*(4), 540–547 (cit. on p. 44).

Kleinschmidt, D. F., Weatherholtz, K. & Florian Jaeger, T. (2018). Sociolinguistic Perception as Inference Under Uncertainty. *Topics in Cognitive Science*, *10*(4), 818–834 (cit. on p. 124).

Kong, E. J. & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, *59*, 40–57 (cit. on p. 126).

Kraljic, T. & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178 (cit. on p. 2).

Kraljic, T. & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268 (cit. on p. 2).

Kuang, J. & Cui, A. (2018). Relative cue weighting in production and perception of an ongoing sound change in Southern Yi. *Journal of Phonetics*, *71*, 194–214 (cit. on p. 93).

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26 (cit. on p. 103).

Labov, W. (1963). The Social Motivation of a Sound Change. *Word*, *19*(3), 273–309 (cit. on pp. 2, 4, 49).

Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics. (Cit. on p. 4).

Labov, W. (1972). The Social Stratification of /r/ in New York City Department Stores. In W. Labov (Ed.), *Sociolinguistic Patterns* (pp. 43–69). University of Pennsylvania Press. (Cit. on pp. 2, 4).

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, *2*(2), 205–254 (cit. on pp. 4, 124).

Labov, W. (1994). *Principles of Linguistic Change. Volume 1: Internal Factors*. Blackwell. (Cit. on p. 16).

Labov, W. (2001). *Principles of Linguistic Change. Volume 2: Social Factors*. Blackwell. (Cit. on pp. 4, 47, 54, 124).

Labov, W. (2007). Transmission and Diffusion. *Language*, *83*(2), 344–387 (cit. on pp. 87, 94).

Labov, W., Baranowski, M. & Dinkin, A. (2010). The effect of outliers on the perception of sound change. *Language Variation and Change*, *22*(2), 175–190 (cit. on p. 122).

Ladd, D. R. (2006). "Distinctive phones" in surface representation. In L. Goldstein, D. H. Whalen & C. T. Best (Eds.), *Laboratory Phonology 8* (pp. 3–26). Mouton de Gruyter. (Cit. on p. 81).

Lakin, J. L. & Chartrand, T. L. (2003). Using Nonconscious Behavioral Mimicry to Create Affiliation and Rapport. *Psychological Science*, *14*(4), 334–339 (cit. on p. 49).

Laurinavichyute, A., Yadav, H. & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language*, *125*, 104332 (cit. on p. 89).

Laver, M. & Sergenti, E. (2011). *Party Competition: An Agent-Based Model*. Princeton University Press. (Cit. on p. 5).

Lee, D. D. & Seung, H. S. (2001). Algorithms for Non-negative Matrix Factorization. *Advances in Neural Information Processing Systems*, *13*, 1–7 (cit. on pp. 59, 192).

Lee, S. & Katz, J. (2016). Perceptual integration of acoustic cues to laryngeal contrasts in Korean fricatives. *The Journal of the Acoustical Society of America*, *139*(2), 605–611 (cit. on p. 126).

Lee, Y., Goldstein, L., Parrell, B. & Byrd, D. (2021). Who converges? Variation reveals individual speaker adaptability. *Speech Communication*, *131*, 23–34 (cit. on p. 49).

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A. & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, *5*(1), 42 (cit. on p. 202).

Lehet, M. & Holt, L. L. (2017). Dimension-Based Statistical Learning Affects Both Speech Perception and Production. *Cognitive Science*, *41*, 885–912 (cit. on pp. 18, 123, 127, 128).

Lenth, R. V. (2022). *Emmeans: Estimated marginal means, aka least-squares means* [R package version 1.7.2]. (Cit. on p. 103).

Lev-Ari, S. (2018). Social network size can influence linguistic malleability and the propagation of linguistic change. *Cognition*, *176*, 31–39 (cit. on pp. 79, 94).

Levelt, W. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, *98*(23), 13464–13471 (cit. on p. 84).

Lin, S., Beddor, P. S. & Coetzee, A. W. (2014). Gestural reduction, lexical frequency, and sound change: A study of post-vocalic /l/. *Laboratory Phonology*, *5*(1), 9–36 (cit. on p. 85).

Lindblom, B. (1988). Phonetic invariance and the adaptive nature of speech. In B. A. G. Elsendoorn & H. Bouma (Eds.), *Working Models of Human Perception* (pp. 139–173). Academic Press. (Cit. on p. 3).

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling*. Springer Netherlands. (Cit. on p. 3).

Lindblom, B. (1998). Systemic constraints and adaptive change in the formation of sound structure. In J. Hurford, M. Studdert-Kennedy & C. Knight (Eds.), *Approaches to the Evolution of Language* (pp. 242–264). Cambridge University Press. (Cit. on p. 3).

Lindblom, B., Guion, S. G., Hura, S. L., Moon, S.-J. & Willerman, R. (1995). Is sound change adaptive? *Rivista di Linguistica*, *7*(1), 5–37 (cit. on pp. 3, 4, 38, 47).

Lisker, L. (1986). "Voicing" in English: A Catalogue of Acoustic Features Signaling /b/ versus /p/ in Trochees. *Language and Speech*, *29*(1), 3–11 (cit. on pp. 93, 128).

Luce, P. A. & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, *19*(1), 1–36 (cit. on pp. 84, 86).

Luthra, S., Guediche, S., Blumstein, S. E. & Myers, E. B. (2019). Neural substrates of subphonemic variation and lexical competition in spoken

word recognition. *Language, Cognition and Neuroscience*, *34*(2), 151–169 (cit. on p. 51).

MacLennan, B. (2007). Evolutionary Psychology, Complex Systems, and Social Theory. *Soundings: An Interdisciplinary Journal*, *90*(3-4), 169–189 (cit. on p. 46).

Mahalanobis, P. C. (1936). *On the generalized distance in statistics* (Proceedings of the National Institute of Sciences of India). National Institute of Sciences of India. Prayagraj. (Cit. on p. 206).

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. (Cit. on p. 199).

Martin, J. G. & Bunnell, H. T. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(3), 473–488 (cit. on p. 18).

Martinet, A. (1952). Function, Structure, and Sound Change. *Word*, *8*(1), 1–32 (cit. on p. 4).

McCullough, E. A., Clopper, C. G. & Wagner, L. (2019). Regional dialect perception across the lifespan: Identification and discrimination. *Language and Speech*, *62*(1), 115–136 (cit. on p. 123).

McMurray, B., Clayards, M. A., Tanenhaus, M. K. & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review*, *15*(6), 1064–1071 (cit. on p. 126).

McMurray, B. & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219–246 (cit. on p. 126).

McQueen, J. M., Norris, D. & Cutler, A. (2006). The Dynamic Nature of Speech Perception. *Language and Speech*, *49*(1), 101–112 (cit. on p. 18).

Mitterer, H., Scharenborg, O. & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356–361 (cit. on p. 51).

Momcilovic, N. B. (2009). *A sociolinguistic analysis of /s/-aspiration in Madrid Spanish*. LINCOM. (Cit. on p. 14).

Mondéjar Cumpián, J. (2001). *Dialectología Andaluza: Estudios. Tomo I* (P. Carrasco & M. Galeote, Eds.). Universidad de Málaga. (Cit. on p. 14).

Moya Corral, J. A. (2007). Noticia de un sonido emergente: La africada dental procedente del grupo -st- en Andalucía. *Revista de Filología de la Universidad de La Laguna*, *25*, 457–466 (cit. on pp. 36, 139).

Munson, B. (2011). Lavender Lessons Learned; Or, What Sexuality Can Teach Us About Phonetic Variation. *American Speech*, *86*(1), 14–31 (cit. on p. 124).

Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142 (cit. on p. 157).

Nardy, A., Chevrot, J.-P. & Barbu, S. (2014). Sociolinguistic convergence and social interactions within a group of preschoolers: A longitudinal study. *Language Variation and Change*, *26*(3), 273–301 (cit. on p. 87).

Navarro Tomás, T. (1938). Dédoublement de phonemes dans le dialecte andalou. *Travaux du Cercle Linguistique de Prague*, *8*, 184–186 (cit. on p. 15).

Néda, Z., Ravasz, E., Brechet, Y., Vicsek, T. & Barabási, A.-L. (2000). The sound of many hands clapping: Tumultuous applause can transform itself into waves of synchronizes clapping. *Nature*, *403*, 849–850 (cit. on p. 49).

Nguyen, N. & Delvaux, V. (2015). Role of imitation in the emergence of phonological systems. *Journal of Phonetics*, *53*, 46–54 (cit. on p. 50).

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142 (cit. on pp. 49, 51).

Nielsen, K. (2014). Phonetic Imitation by Young Children and Its Developmental Changes. *Journal of Speech Language and Hearing Research*, *57*, 2065–2075 (cit. on p. 87).

Norris, D., McQueen, J. M. & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238 (cit. on pp. x, xv, 2, 18, 50).

Oden, G. C. & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*(3), 172–191 (cit. on p. 126).

Ohala, J. J. (1981). The Listener as a Source of Sound Change. In C. S. Masek, R. A. Hendrick & M. F. Miller (Eds.), *Papers from the Parasession on Language and Behavior* (pp. 178–203). Chicago Linguistic Society. (Cit. on pp. 3, 127).

Ohala, J. J. (1983). The Origin of Sound Patterns in Vocal Tract Constraints. In P. F. MacNeilage (Ed.), *The Production of Speech* (pp. 189–216). Springer New York. (Cit. on pp. 2, 49).

Ohala, J. J. (1989). Sound change is drawn from a pool of synchronic variation. In L. E. Breivik & E. H. Jahr (Eds.), *Language Change: Contributions to the study of its causes* (pp. 173–198). Mouton de Gruyter. (Cit. on pp. 2, 3, 47, 49).

Ohala, J. J. (1990). The phonetics and phonology of aspects of assimilation. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 258–282). Cambridge University Press. (Cit. on pp. 16, 93).

Ohala, J. J. (1993a). The phonetics of sound change. In C. Jones (Ed.), *Historical Linguistics: Problems and Perspectives* (pp. 237–278). Longman. (Cit. on pp. 3, 38).

Ohala, J. J. (1993b). Sound change as nature's speech perception experiment. *Speech Communication*, *13*(1-2), 155–161 (cit. on pp. 3, 14).

Ohala, J. J. (1997). Aerodynamics of phonology. *Proceedings of the 4th Seoul International Conference on Linguistics*, 92–97 (cit. on pp. 2, 49).

Ohala, J. J. (2012). The listener as a source of sound change: An update. In M.-J. Solé & D. Recasens (Eds.), *The Initiation of Sound Change: Perception, Production, and Social Factors* (pp. 21–36). John Benjamins. (Cit. on pp. 3, 47, 127).

Ohala, J. J. & Amador, M. (1981). Spontaneous nasalization. *The Journal of the Acoustical Society of America*, *69*(S1), S54–S55 (cit. on p. 125).

Ohala, J. J. & Ohala, M. (1993). The phonetics of nasal phonology: Theorems and data. In M. Huffman & R. A. Krakow (Eds.), *Nasals, Nasalization, and the Velum* (pp. 225–249). Academic Press. (Cit. on p. 92).

Oliveira, L. & Marin, S. (2005). Patterns of velum coordination in Brazilian Portuguese. *Proceedings of Phonetics and Phonology in Iberia*, 1–2 (cit. on p. 93).

O'Neill, P. (2009). The effect of s-aspiration on occlusives in Andalusian Spanish. *Proceedings of the Oxford University Working Papers in Linguistics, Philology, and Phonetics*, 73–86 (cit. on p. 109).

O'Neill, P. (2010). Variación y cambio en las consonantes oclusivas del español de Andalucía. *Estudios de fonética experimental*, *19*, 11–41 (cit. on pp. 36–38).

Ou, J., Yu, A. C. L. & Xiang, M. (2021). Individual Differences in Categorization Gradience As Predicted by Online Processing of Phonetic Cues During Spoken Word Recognition: Evidence From Eye Movements. *Cognitive Science*, *45*(3), e12948 (cit. on p. 126).

Palmeri, T. J., Goldinger, S. D. & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328 (cit. on pp. 2, 50).

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393 (cit. on pp. 49, 84).

Pardo, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, *4*, 1–5 (cit. on p. 49).

Pardo, J. S., Gibbons, R., Suppes, A. & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190–197 (cit. on pp. x, xv, 49).

Parrell, B. (2012). The role of gestural phasing in Western Andalusian Spanish aspiration. *Journal of Phonetics*, *40*(1), 37–45 (cit. on pp. vii, xiii, 6, 15–17, 19, 20, 38, 92, 108, 109, 121).

Pascual-Montano, A., Carazo, J.-M., Kochi, K., Lehmann, D. & Pascual-Marqui, R. D. (2006). Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(3), 403–415 (cit. on p. 193).

Penzl, H. (1949). Umlaut and Secondary Umlaut in Old High German. *Language*, *25*(3), 223–240 (cit. on p. 19).

Peterson, G. E. & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184 (cit. on p. 2).

Phillips, B. S. (1984). Word Frequency and the Actuation of Sound Change. *Language*, *60*(2), 320–342 (cit. on pp. 4, 85, 86).

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. J. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (pp. 137–155). John Benjamins. (Cit. on pp. ix, xv, 2, 50, 51, 67, 80, 81, 84, 85, 96, 124).

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 101–140). De Gruyter Mouton. (Cit. on pp. 84, 124).

Pierrehumbert, J. B. (2003a). Phonetic Diversity, Statistical Learning, and Acquisition of Phonology. *Language and Speech*, *46*(2-3), 115–154 (cit. on pp. ix, xv, 2, 39, 51, 81, 123).

Pierrehumbert, J. B. (2003b). Probabilistic Phonology: Discrimination and Robustness. In R. Bod, J. B. Hay & S. Jannedy (Eds.), *Probability Theory in Linguistics* (pp. 177–228). MIT Press. (Cit. on p. 51).

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics*, *34*(4), 516–530 (cit. on pp. 2, 39).

Pierrehumbert, J. B., Stonedahl, F. & Daland, R. (2014). A model of grassroots changes in linguistic systems. *Computing Research Repository*, *abs / 1408.1985*, 1–30 (cit. on pp. 77, 94).

Pompino-Marschall, B. (2009). *Einführung in die Phonetik* (3. ed.). De Gruyter. (Cit. on p. 1).

Purnell, T., Idsardi, W. & Baugh, J. (1999). Perceptual and Phonetic Experiments on American English Dialect Identification. *Journal of Language and Social Psychology*, *18*, 10–30 (cit. on p. 123).

Ramsammy, M. (2015). The Life Cycle of Phonological Processes: Accounting for Dialectal Microtypologies. *Language and Linguistics Compass*, *9*(1), 33–54 (cit. on pp. 19, 139).

Ramsay, J., Graves, S. & Hooker, G. (2021). *Fda: Functional data analysis* [R package version 5.5.1]. (Cit. on p. 99).

Ramsay, J. O. & Silverman, B. W. (2010). *Functional Data Analysis*. Springer New York. (Cit. on pp. 24, 99).

Reinisch, E. & Holt, L. L. (2013). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 539–555 (cit. on p. 18).

Reinisch, E., Juhl, K. I. & Llompart, M. (2020). The Impact of Free Allophonic Variation on the Perception of Second Language Phonological Categories. *Frontiers in Communication*, *5*, 47 (cit. on p. 51).

Reinisch, E. & Mitterer, H. (2016). Exposure modality, input variability and the categories of perceptual recalibration. *Journal of Phonetics*, *55*, 96–108 (cit. on p. 51).

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*(1), 81–110 (cit. on pp. 18, 93).

Reubold, U. & Harrington, J. (2015). Disassociating the effects of age from phonetic change: A longitudinal study of formant frequencies. In A. Gerstenberg & A. Voeste (Eds.), *Language Development: The Lifespan Perspective* (pp. 9–37). John Benjamins. (Cit. on p. 85).

Reynolds, D. (2009). Gaussian Mixture Models. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of Biometrics* (pp. 659–663). Springer US. (Cit. on pp. 59, 190).

Richardson, M. J., Marsh, K. L. & Schmidt, R. C. (2005). Effects of Visual and Verbal Interaction on Unintentional Interpersonal Coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(1), 62–79 (cit. on p. 49).

Roberts, J. (1997). Hitting a moving target: Acquisition of sound change in progress by Philadelphia children. *Language Variation and Change*, *9*, 249–266 (cit. on p. 87).

Roberts, J. & Labov, W. (1995). Learning to talk Philadelphian: Acquisition of short a by preschool children. *Language Variation and Change*, *7*, 101–112 (cit. on p. 87).

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*, 107–142 (cit. on p. 84).

Roettger, T. B., Winter, B. & Baayen, H. (2019). Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics*, *73*, 1–7 (cit. on p. 89).

Romero, J. (1994). An Articulatory View of Historical S-aspiration in Spanish. *Haskins Laboratories Status Report on Speech Research*, 255–266 (cit. on pp. 14, 15).

Ruch, H. (2010). Affrication of /st/-clusters in Western Andalusian Spanish: Variation and change from a sociophonetic point of view. *Proceedings of the Workshop 'Sociophonetics, at the Crossroads of Speech Variation, Processing and Communication'*, 61–64 (cit. on pp. 139, 141).

Ruch, H. (2013). *Lautvariation und Lautwandel im andalusischen Spanisch: Prä- und Postaspiration bei /s/ vor stimmlosen Plosiven* (Doctoral dissertation). LMU. München. (Cit. on pp. 16, 20, 21, 36, 42, 92, 97, 98, 108, 151).

Ruch, H. (2018). Perception of speaker age and speaker origin in a sound change in progress: The case of /s/-aspiration in Andalusian Spanish. *Journal of Linguistic Geography*, *6*(1), 40–55 (cit. on pp. 16, 92, 93, 123).

Ruch, H. & Harrington, J. (2014). Synchronic and diachronic factors in the change from pre-aspiration to post-aspiration in Andalusian Spanish. *Journal of Phonetics*, *45*, 12–25 (cit. on pp. vii, xiii, 14, 16, 18, 20–22, 34, 38, 41, 42, 92, 93, 108, 151, 161).

Ruch, H. & Peters, S. (2016). On the Origin of Post-Aspirated Stops: Production and Perception of /s/ + Voiceless Stop Sequences in Andalusian Spanish. *Laboratory Phonology*, *7*(1), 1–36 (cit. on pp. 15, 16, 20, 21, 34, 37, 38, 41, 42, 139, 151).

Salmons, J. C., Fox, R. A. & Jacewicz, E. (2012). Prosodic skewing of input and the initiation of cross-generational sound change. In M.-J. Solé & D. Recasens (Eds.), *The Initiation of Sound Change: Perception, Production, and Social Factors* (pp. 167–184). John Benjamins. (Cit. on p. 89).

Saltzman, D. & Myers, E. (2021). Listeners are initially flexible in updating phonetic beliefs over time. *Psychonomic Bulletin & Review*, *28*(4), 1354–1364 (cit. on p. 50).

Samuel, A. G. & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218 (cit. on pp. 2, 18, 50).

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J.-L. & Nguyen, N. (2013). Converging toward a common speech code: Imitative and perceptuo-motor recalibration processes in speech production. *Frontiers in Psychology*, *4*, 1–14 (cit. on p. 49).

Savoia, L. & Maiden, M. (1997). Metaphony. In M. Maiden (Ed.), *The Dialects of Italy* (pp. 15–25). Routledge. (Cit. on p. 18).

Schelling, T. C. (1978). *Micromotives and macrobehavior*. Norton. (Cit. on pp. ix, xv, 46).

Schertz, J., Cho, T., Lotto, A. & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183–204 (cit. on pp. 126, 128).

Schertz, J. & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *WIREs Cognitive Science*, *11*(2), e1521 (cit. on pp. 93, 94, 128).

Schiel, F., Draxler, C. & Harrington, J. (2011). Phonemic segmentation and labelling using the MAUS technique. *Proceedings of the Workshop New Tools and Methods for Very-Large-Scale Phonetics Research* (cit. on p. 98).

Schlüter, J. C., Sörensen, L., Bossert, A., Kersting, M., Staab, W. & Wacker, B. (2021). Anticipating the impact of COVID19 and comorbidities on the South African healthcare system by agent-based simulations. *Scientific Reports*, *11*(1), 7901 (cit. on p. 5).

Schubiger, M. (1970). *Einführung in die Phonetik*. De Gruyter. (Cit. on p. 1).

Scobbie, J. M. (2006). Flexibility in the face of incompatible English VOT systems. In L. Goldstein, D. H. Whalen & C. T. Best (Eds.), *Laboratory Phonology 8* (pp. 367–392). Mouton de Gruyter. (Cit. on p. 82).

Scobbie, J. M. & Stuart-Smith, J. (2008). Quasi-phonemic contrast and the fuzzy inventory: Examples from Scottish English. In P. Avery, B. E. Dresher & K. Rice (Eds.), *Contrast in Phonology: Theory, Perception, Acquisition* (pp. 87–113). Mouton de Gruyter. (Cit. on pp. 51, 82).

Scrucca, L., Fop, M., Murphy, B. T. & Raftery, A. E. (2016). Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, *8*(1), 289 (cit. on p. 191).

Sebanz, N., Bekkering, H. & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76 (cit. on p. 50).

Shahnaz, F., Berry, M. W., Pauca, V. P. & Plemmons, R. J. (2006). Document clustering using nonnegative matrix factorization. *Information Processing & Management*, *42*(2), 373–386 (cit. on p. 192).

Sheffert, S. M. & Fowler, C. A. (1995). The Effects of Voice and Visible Speaker Change on Memory for Spoken Words. *Journal of Memory and Language*, *34*, 665–685 (cit. on p. 2).

Shockley, K., Sabadini, L. & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422–429 (cit. on p. 49).

Silverman, D. (2003). On the rarity of pre-aspirated stops. *Journal of Linguistics*, *39*(3), 575–598 (cit. on p. 6).

Smith, B. J., Mielke, J., Magloughlin, L. & Wilbanks, E. (2019). Sound change and coarticulatory variability involving English /ɹ/. *Glossa*, *4*(1), 1–51 (cit. on p. 87).

Solé, M.-J. (1995). Spatio-Temporal Patterns of Velopharyngeal Action in Phonetic and Phonological Nasalization. *Language and Speech*, *38*(1), 1–23 (cit. on p. 125).

Solé, M.-J. (2010). Effects of syllable position on sound change: An aerodynamic study of final fricative weakening. *Journal of Phonetics*, *38*(2), 289–305 (cit. on pp. 4, 14).

Sóskuthy, M. (2013). *Phonetic Biases and Systemic Effects in the Actuation of Sound Change* (Doctoral dissertation). University of Edinburgh. Edinburgh. (Cit. on pp. 2, 4).

Sóskuthy, M. (2015). Understanding change through stability: A computational study of sound change actuation. *Lingua*, *163*, 40–60 (cit. on pp. x, xvi, 48, 77–79, 82, 83, 88, 94, 125).

Sóskuthy, M., Foulkes, P., Hughes, V. & Haddican, B. (2018). Changing Words and Sounds: The Roles of Different Cognitive Units in Sound Change. *Topics in Cognitive Science*, *10*(4), 787–802 (cit. on p. 4).

Stanford, J. N. & Kenny, L. A. (2013). Revisiting transmission and diffusion: An agent-based model of vowel chain shifts across large communities. *Language Variation and Change*, *25*(2), 119–153 (cit. on pp. 77, 78, 85–88, 94).

Stevens, M. & Harrington, J. (2014). The Individual and the Actuation of Sound Change. *Loquens*, *1*(1), 1–15 (cit. on pp. 4, 47, 48, 78).

Stevens, M., Harrington, J. & Schiel, F. (2019). Associating the origin and spread of sound change using agent-based modelling applied to /s/-retraction in English. *Glossa*, *4*(1), 1–30 (cit. on pp. 39, 73, 80, 81, 88, 123, 135).

Sumner, M., Kim, S. K., King, E. & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology*, *4*, 1015 (cit. on p. 124).

Tamminga, M. (2013). Sound Change without Frequency Effects: Ramifications for Phonological Theory. *Proceedings of the 31st West Coast Conference on Formal Linguistics*, 457–465 (cit. on p. 85).

Terrell, T. D. (1980). Diachronic reconstruction by dialect comparison of variable constraints: -s aspiration and deletion in Spanish. In D. Sankoff & H. Cedergren (Eds.), *Variation omnibus* (pp. 115–124). Linguistic Research. (Cit. on pp. 14–16, 121).

Todd, S., Pierrehumbert, J. B. & Hay, J. B. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, *185*, 1–20 (cit. on pp. x, xvi, 39, 56, 57, 74, 77, 79, 82, 85, 86, 94, 128).

Tonn, G. L. & Guikema, S. D. (2018). An Agent-Based Model of Evolving Community Flood Risk. *Risk Analysis*, *38*(6), 1258–1278 (cit. on p. 5).

Torreira, F. (2006). Coarticulation between Aspirated-s and Voiceless Stops in Spanish: An Interdialectal Comparison. *Selected Proceedings of the 9th Hispanic Symposium*, 113–120 (cit. on pp. vii, xiii, 14–16, 36, 108).

Torreira, F. (2007). Pre- and postaspirated stops in Andalusian Spanish. In P. Prieto, J. Mascaró & M.-J. Solé (Eds.), *Prosodic and Segmental Issues in*

Romance (pp. 67–82). John Benjamins. (Cit. on pp. 14, 15, 20, 36, 42, 92, 108, 151).

Torreira, F. (2012). Investigating the nature of aspirated stops in Western Andalusian Spanish. *Journal of the International Phonetic Association*, *42*(1), 49–63 (cit. on pp. vii, xiii, 16, 92).

Torres-Tamarit, F., Linke, K. & van Oostendorp, M. (Eds.). (2016). *Approaches to metaphony in the languages of Italy*. De Gruyter Mouton. (Cit. on pp. 18, 125).

Toscano, J. C. & McMurray, B. (2010). Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics. *Cognitive Science*, *34*(3), 434–464 (cit. on pp. 94, 126).

Toscano, J. C. & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284–1301 (cit. on pp. 94, 126).

Trudgill, P. (1986). *Dialects in Contact*. Blackwell. (Cit. on p. 4).

Trudgill, P. (1999). Dialect contact, dialectology and sociolinguistics. *Cuadernos de Filología Inglesa*, *8*, 1–8 (cit. on pp. 4, 49).

Trudgill, P. (2004). *New-Dialect Formation: The Inevitability of Colonial Englishes*. Oxford University Press. (Cit. on pp. 4, 47, 87).

Trudgill, P. (2008a). Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation. *Language in Society*, *37*, 241–280 (cit. on pp. 4, 49).

Trudgill, P. (2008b). On the role of children, and the mechanical view: A rejoinder. *Language in Society*, *37*(2), 277–280 (cit. on pp. 4, 87).

Trudgill, P. (2011). Contact and isolation in phonology. In P. Trudgill (Ed.), *Sociolinguistic Typology: Social Determinants of Linguistic Complexity* (pp. 116–145). Oxford University Press. (Cit. on p. 4).

Trudgill, P., Gordon, E., Lewis, G. & Maclagan, M. (2000). Determinism in new-dialect formation and the genesis of New Zealand English. *Journal of Linguistics*, *36*, 299–318 (cit. on pp. 4, 50, 53).

Tuller, B. & Kelso, J. A. S. (1989). Phase transitions in speech production and their perceptual consequences. *The Journal of the Acoustical Society of America*, *86*, 114 (cit. on p. 93).

Twaddell, W. F. (1938). A Note on Old High German Umlaut. *Monatshefte für Deutschen Unterricht*, *30*(3/4), 177–181 (cit. on pp. 19, 93).

Uehara, S. & Evans Wagner, S. (2017). Progressive outliers in listener perception of sound change. *Proceedings of the Conference on New Ways of Analyzing Variation*, 1–2 (cit. on p. 122).

van der Kooij, E. & van der Hulst, H. (2005). On the internal and external organization of sign language segments: Some modality-specific properties. In M. van Oostendorp & J. van de Weijer (Eds.), *The Internal Organization of Phonological Segments* (pp. 153–180). Mouton de Gruyter. (Cit. on p. 20).

Vida Castro, M. (2016). Correlatos acústicos y factores sociales en la aspiración de /-s/ preoclusiva en la variedad de Málaga (España). Análisis de un cambio fonético en curso. *Lingua Americana*, *38*, 15–36 (cit. on p. 141).

Villena-Ponsoda, J. A. (2008). Sociolinguistic patterns of Andalusian Spanish. *International Journal of the Sociology of Language*, *193/194*, 139–160 (cit. on pp. vii, xiii, 14, 15, 92).

Vitevitch, M. S. & Luce, P. A. (1999). Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *Journal of Memory and Language*, *40*(3), 374–408 (cit. on p. 86).

Wedel, A. (2006). Exemplar models, evolution and language change. *The Linguistic Review*, *23*(3), 247–274 (cit. on pp. 39, 77, 78, 82–84, 94, 128).

Weinreich, U., Labov, W. & Herzog, M. I. (1968). Empirical Foundations for a Theory of Language Change. In W. P. Lehmann & Y. Malkiel (Eds.), *Directions for Historical Linguistics* (pp. 95–195). University of Texas Press. (Cit. on pp. 3, 4, 48, 125, 129).

Whalen, D. H., Abramson, A. S., Lisker, L. & Mody, M. (1990). Gradient Effects of Fundamental Frequency on Stop Consonant Voicing Judgments. *Phonetica*, *47*, 36–49 (cit. on p. 18).

Wimmers, R. H., Beek, P. J. & van Wieringen, P. C. (1992). Phase transitions in rhythmic tracking movements: A case of unilateral coupling. *Human Movement Science*, *11*(1-2), 217–226 (cit. on p. 93).

Winkelmann, R. (2017). *The EMU-SDMS* (Doctoral dissertation). LMU Munich. Munich. (Cit. on p. 98).

Winkler, I., Paavilainen, P., Alho, K., Reinikainen, K., Sams, M. & Naatanen, R. (1990). The Effect of Small Variation of the Frequent Auditory Stimulus on the Event-Related Brain Potential to the Infrequent Stimulus. *Psychophysiology*, *27*(2), 228–235 (cit. on p. 123).

Winter, B. (2020). *Statistics for Linguists: An Introduction Using R* (1st ed.). Routledge. (Cit. on p. 89).

Winter, B. & Wedel, A. (2016). The Co-evolution of Speech and the Lexicon: The Interaction of Functional Pressures, Redundancy, and Category Variation. *Topics in Cognitive Science*, *8*(2), 503–513 (cit. on p. 128).

Wise, S. C. & Cheng, T. (2016). How Officers Create Guardianship: An Agent-based Model of Policing. *Transactions in GIS*, *20*(5), 790–806 (cit. on p. 5).

Wong, P. & Babel, M. (2017). Perceptual identification of talker ethnicity in Vancouver English. *Journal of Sociolinguistics*, *21*(5), 603–628 (cit. on p. 123).

Yu, A. C. L. (2013). Individual differences in socio-cognitive processing and the actuation of sound change. In A. C. L. Yu (Ed.), *Origins of Sound Change* (pp. 201–227). Oxford University Press. (Cit. on pp. 2, 4, 48).

Yu, A. C. L. (2021). Toward an individual-difference perspective on phonologization. *Glossa*, *6*(1), 1–24 (cit. on pp. 78, 123, 126).

Yu, A. C. L. (2022). Perceptual Cue Weighting Is Influenced by the Listener's Gender and Subjective Evaluations of the Speaker: The Case of English Stop Voicing. *Frontiers in Psychology*, *13*, 840291 (cit. on p. 126).

Yu, A. C. L. & Zellou, G. (2019). Individual Differences in Language Processing: Phonology. *Annual Review of Linguistics*, *5*(1), 131–150 (cit. on pp. 2, 78).

Zellou, G. (2017). Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics*, *61*, 13–29 (cit. on pp. 2, 42).

Zellou, G. & Brotherton, C. (2021). Phonetic imitation of multidimensional acoustic variation of the nasal split short-a system. *Speech Communication*, *135*, 54–65 (cit. on p. 49).

Zellou, G. & Tamminga, M. (2014). Nasal coarticulation changes over time in Philadelphia English. *Journal of Phonetics*, *47*, 18–35 (cit. on p. 92).

Zhang, X. & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 200–217 (cit. on p. 2).

List of Figures

1.1 Waveform and spectrogram of the word *despide* by two Andalusian Spanish speakers . . . 7

2.1 Idealised scheme of gestural resynchronisation in Andalusian Spanish . . . 17
2.2 Example of high-frequency energy and voicing probability signals 25
2.3 Example of analysis for a post- and a pre-aspirated token . . . 28
2.4 Variation expressed by PC1 . . . 30
2.5 Aspiration areas as a function of PC score s_1 . . . 32
2.6 Boxplots of s_1 values by speaker age, region, and cluster . . . 36
2.7 Reconstruction of HF and VP using PC1 by age, region, and cluster . . . 37
2.8 Outline of association between FPCA outputs and a model of sound change . . . 40
2.9 Post- against pre-aspiration area . . . 43

3.1 Schematic sketch of the three focal points of the IP model . . . 48
3.2 Example of two agents' phonological classes and their relation to the memorisation criteria . . . 57
3.3 Example of GMM and NMF . . . 60
3.4 Results of GMM and NMF in case of systematic data . . . 66
3.5 Results of GMM and NMF in case of unsystematic data . . . 67
3.6 Results of a simulation that shows a phonetic shift . . . 70
3.7 Exemplars rejected or accepted by the absolute memorisation criterion . . . 71
3.8 Results of a simulation that shows phoneme repulsion . . . 72
3.9 Exemplars rejected or accepted by the relative memorisation criterion . . . 74
3.10 Results of a simulation with both memorisation criteria . . . 75

3.11 Exemplars accepted by both memorisation criteria or rejected by either of them 77

4.1 Variation expressed by PC1, PC2, PC3, and PC4 101
4.2 PC scores s_1, s_2, s_3, and s_4 for /st/ and /t/ by age and region. . 104
4.3 Reconstructed HF and VP for /st/ and /t/ by age and region . . 107
4.4 Reconstructed HF and VP separately for age groups, plosives, and simulation state 115
4.5 Estimated marginal means with 95% confidence bars for s_1, s_2, and s_4 116
4.6 Number of sub-phonemes and canonical agreement over simulated time 119

5.1 Waveform and spectrogram of the word *pasta* by two Andalusian Spanish speakers 140

A.1 Aspiration areas as a function of s_1 147
A.2 Random partition of the areas between HF and VP 149
A.3 Empirical Cumulative Distribution Function (ECDF) of P'_{pre} . . 150
A.4 Post- against pre-aspiration areas, coloured by s_1 151
A.5 Examples of reconstructed post-aspirated tokens which differ in s_2 153
A.6 Variation captured by PC2 and PC3 154
A.7 Boxplots of s_2 and s_3 by speaker age, region, and cluster 155
A.8 Variation expressed by PC1, including duration 159
A.9 Aspiration areas as a function of s_1 160
A.10 Distribution of pre- and post-aspiration measures obtained using different methods 164

B.1 General scheme of the implementation of acoustic and sub-phonemic representations in `soundChangeR`. 196
B.2 Example of how acoustic clusters and sub-phonemic classes are identified using GMM and NMF 197
B.3 Example of application of SMOTE to a set of four points 203

B.4 Schematic representation of a token exchange between an agent speaker and an agent listener 206

List of Tables

3.1 Association between words and classical phonemes 61

A.1 Pseudo-R^2 for the LMERs 157
A.2 Specifications for scatter plots in Figure A.10. 161

B.1 Symbols used in Appendices B.4.3, B.5.2, and B.6 195
B.2 Exemplar counts by word type for acoustic clusters, sub-phonemic clusters, and assignment of word types to sub-phonemic classes 200
B.3 Five purity estimates and their average values for different amounts of sub-phonemic clusters 201

C.1 Median PC scores used in Figure 4.3 212
C.2 Median PC scores used in Figure 4.4 212
C.3 Estimated marginal means for s_1 213
C.4 Estimated marginal means for s_2 213
C.5 Estimated marginal means for s_3 214
C.6 Estimated marginal means for s_4 214