https://doi.org/10.35662/unine-thesis-3204

UNIVERSITY OF NEUCHÂTEL

DOCTORAL THESIS

Morpho-phonetics: A Framework for Studying Speech and Its Evolution

Author:
Tiena DANNER

Supervisor:
Prof. Dr. Steven MORAN

Co-Advisors:
Prof. Dr. Christoph ZOLLIKOFER
Dr. Marcia PONCE DE LEÓN

Jury:
Prof. Dr. Steven MORAN,
Prof. Dr. Klaus ZUBERBÜHLER,
Dr. François PELLEGRINO,
Prof. Dr. Dan DEDIU

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Laboratory of Language Evolution
Institute of Biology

unine
Université de Neuchâtel

Date of defense: 04.07.2025

Declaration of Authorship

I, Tiena DANNER, declare that this thesis titled, "Morpho-phonetics: A Framework for Studying Speech and Its Evolution" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help (see below).
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- In this work I used ChatGPT (Open AI, https://chat.openai.com/chat) to assist with grammatical corrections and to improve the clarity and style of my own writing.

Signed:

Date:

"*All that is gold does not glitter,*
Not all those who wander are lost."

– J. R. R. Tolkien, The Lord of the Rings: The Fellowship of the Ring

UNIVERSITY OF NEUCHÂTEL

Abstract

Faculty of Science
Institute of Biology

Doctor of Philosophy

Morpho-phonetics: A Framework for Studying Speech and Its Evolution

by Tiena DANNER

Linguistic diversity arises from a complex interplay of biological, cultural, and environmental forces that shape the acoustic, phonological, lexical, and grammatical systems of the world's languages. In this dissertation I explore how acoustic diversity propagates from interindividual variation in vocal tract morphology to systematic acoustic differences across populations and languages. Focusing on vowel systems, I develop and apply a novel framework termed *morpho-phonetics* – which integrates geometric morphometric analysis of speech articulation with real-time MRI and synchronized acoustic analysis – to examine how anatomical differences in the vocal tract give rise to distinct articulatory strategies and consistent acoustic signatures. These insights extend to speech phenomena such as coarticulation, offering new tools to investigate continuous variation in connected speech and showing that phonetic contrast is maintained through relative, rather than absolute, articulatory positioning.

Scaling from individuals to populations, I then apply evolutionary models to large-scale speech datasets, uncovering isolation-by-distance (IBD) patterns in vowel acoustics that mirror drift-like processes of gradual divergence. This supports a model of neutral-like evolution of acoustics in spoken language, where geographic distance influences acoustic differentiation. Importantly, these dynamics are not limited to humans: parallel analyses of Eurasian chaffinch birdsong reveal strikingly similar IBD and neutral-like patterns, suggesting that both speech and birdsong may be shaped by comparable population-level processes. These findings provide a biologically grounded null model of acoustic trait evolution, against which non-neutral or culturally driven changes can be tested.

Finally, I demonstrate how acoustic features may retain historical signals over deep timescales, potentially exceeding the 8–10k-year limit of the comparative method. Similarity-based clustering of vowel acoustics recovers known genealogical relationships between language families, indicating that gradual phonetic divergence may serve as a powerful, underutilized tool in historical linguistics. This work proposes a new approach to language diversification rooted in acoustic analysis, offering insights into both the evolution of language and the mechanisms that maintain its diversity over time.

Keywords: geometric morphometrics, real-time MRI, coarticulation, language evolution, neutral evolution

Résumé de la thèse

La diversité linguistique résulte d'une interaction complexe de forces biologiques, culturelles et environnementales qui façonnent les systèmes acoustiques, phonologiques, lexicaux et grammaticaux des langues du monde. Cette thèse explore la façon dont cette diversité se propage de la variation interindividuelle de la morphologie du conduit vocal aux différences acoustiques systématiques entre les populations et les langues.

En me concentrant sur les systèmes vocaliques, je développe et applique un nouveau modèle appelé *morpho-phonétique* – intégrant l'analyse morphométrique géométrique de l'articulation de la parole avec un IRM en temps réel et l'analyse acoustique synchronisée – pour examiner comment les différences anatomiques dans le conduit vocal donnent lieu à des stratégies articulatoires distinctes et à des signatures acoustiques cohérentes. Ces connaissances s'étendent à des phénomènes vocaux tels que la coarticulation, offrant de nouveaux outils pour étudier la variation continue dans la parole connectée et montrant que le contraste phonétique est maintenu par un positionnement articulatoire relatif plutôt qu'absolu.

En passant des individus aux populations, j'applique ensuite des modèles évolutifs à des ensembles de données vocales à grande échelle, découvrant des schémas d'isolement par la distance (IBD) dans l'acoustique des voyelles qui reflètent des processus de dérive semblables à une divergence phonétique graduelle. Ceci soutient un modèle d'évolution de type neutre de l'acoustique dans la langue parlée, où la distance géographique influence la différenciation acoustique. Il est important de noter que cette dynamique ne se limite pas à l'homme: des analyses parallèles du chant des oiseaux du pinson d'Eurasie révèlent des modèles IBD et neutres étonnamment similaires, ce qui suggère que la parole et le chant des oiseaux peuvent être façonnés par des processus comparables au niveau populationnel. Ces résultats fournissent un modèle nul biologiquement fondé de l'évolution des traits acoustiques, qui permet aux changements non neutres ou d'origine culturelle d'être testés.

Enfin, je démontre que les caractéristiques acoustiques peuvent conserver des signaux historiques sur des échelles de temps très longues, dépassant potentiellement la limite de 8,000 à 10,000 ans de la méthode comparative. Le regroupement de l'acoustique des voyelles basé sur la similarité permet de retrouver des relations généalogiques connues entre les familles de langues, ce qui indique que la divergence phonétique graduelle peut constituer un outil puissant et sous-utilisé en linguistique historique. Ce travail propose une nouvelle approche de la diversification des langues fondée sur l'analyse acoustique, qui permet de mieux comprendre à la fois l'évolution des langues et les mécanismes qui maintiennent leur diversité au fil du temps.

Mots-clés: morphométrie géométrique, IRM en temps réel, coarticulation, évolution du langage, évolution neutre

Acknowledgements

"No one can whistle a symphony. It takes an orchestra to play it." — H. E. Luccock

This dissertation resembles in many ways a symphony – perhaps not in flawless harmony or elegance (I will leave that for the reader to decide) – but in the spirit of being a truly interdisciplinary and collaborative endeavor. It spans a wide spectrum of disciplines, from anatomy, geometric morphometrics, and evolutionary biology to acoustic phonetics, linguistics, and language evolution. Just as a symphony requires a diverse ensemble of instrumentalists to bring it to life, this dissertation was shaped by the insights, guidance, and camaraderie of brilliant scientists, mentors, and friends. Their collective wisdom helped me explore new territories of thought and discovery. Without these companions on my journey, this work would never have reached its current form. Thank you for helping me compose this symphony. And now, to the most important instrumentalists and conductors of this piece...

First and foremost, I want to thank Steven, my supervisor and good friend, who supported me throughout the whole journey with unwavering patience, goodwill, and scientific brilliance. Even through valleys of struggle, impostor syndrome and self-doubt he carried me with ease. To say the least, you gave me a wonderful introduction and opened the door to the beautiful and interesting world of linguistics. Thank you for letting me stand on your shoulders, such that I could see further. For this, and for your friendship, I will be forever grateful.

My heartfelt thanks also go to Marcia and Christoph, who have been supporting and guiding me since the beginning of my Master's thesis. Your scientific insight, sense for the bigger ideas and broader picture, as well as your remarkable patience and vast knowledge astonishes me to this day. Thank you for all your counseling, for your brilliant ideas, and for your most valuable feedback. I will never forget the most interesting scientific discussions and brain-storming sessions that I had together with you and Steven over IPA (you know what I mean – to many more!).

Further thanks are due to all those who contributed to the collection of data – without their efforts, this dissertation would not have been possible. I extend my sincere gratitude to the team at the *Speech Production and Articulation Knowledge (SPAN)* group at the University of Southern California for generously making their MRI and acoustic data openly available. I am also deeply thankful to the creators of the *VoxCommunis* database for providing open-access, cross-linguistic vowel formant measurements. And of course to Daniel Friedrichs, who helped with collecting more cross-linguistic acoustic data.

I also want to thank the birds, who provided much-needed distractions during moments of stress and writer's block. You continue to enlighten and enrich my life in ways I could never have anticipated, no matter how many times I've watched you. This work is also dedicated to you. I promise to use all that is within my power to protect you and the natural world you embody.

Lastly, and most importantly, I want to thank my closest friends and family, foremost my partner Aline and my daughter Lilly (who was born throughout the finalizing of this work), who have been of great support when I was stressing and especially during the last few weeks of finishing this dissertation, when it got turbulent at times. Thank you for your love and patience.

Contents

List of Figures

List of Tables

Dedicated to my dearly loved Aline and Lilly, and to Science –
my eternal lights in dark places, when all other lights go out.

Introduction

1.1 Rationale

It is a remarkable feature of human nature that, regardless of our genetic background or the environment in which we are raised, children have the capacity to learn any of the world's languages (Chomsky, 2005). Currently, around 8,000 spoken languages exist across the globe (Hammarström et al., 2024). While signed languages are equally complex, expressive, and governed by rich linguistic structure (Sandler and Lillo-Martin, 2006), this work focuses specifically on spoken languages. What produces this vast diversity – so extensive that it often renders us mutually unintelligible – remains one of the most intriguing and yet not fully resolved questions in the study of language (Hauser et al., 2014).

Although we speak different languages, spoken languages are ultimately grounded in speech sounds: vowels, consonants, and tones. Many of these sounds are shared across languages; for instance, the vowel [i] and the consonant [m] are found in over 90% of the world's languages (Moran and McCloy, 2019). These speech sounds are produced by our vocal tracts, which themselves vary considerably across individuals (Weirich and Fuchs, 2011). This anatomical variation is apparent in our daily lives: we can easily identify family members or close friends by their voice alone (Lavan and McGettigan, 2023).

This raises a compelling question: how does interspeaker variation in the anatomy of the speech organs influence the production of speech sounds – and, moreover, the structure and diversity of languages? Speech sounds form the acoustic foundation of language (Ladefoged and Maddieson, 1996), and may encapsulate evolutionary signals that reflect the processes shaping linguistic diversity (Dediu et al., 2019; Moran et al., 2024). In parallel, the cultural transmission of language – through interaction, learning, and adaptation – continuously shapes and amplifies this diversity across generations and populations (Kirby et al., 2008). Understanding how variation in human vocal anatomy affects spoken language production could therefore provide important insights into the broader question of language evolution. Recent advances in imaging technologies, computational modeling, and cross-linguistic phonetic databases now make it increasingly feasible to study these relationships in fine-grained detail (Narayanan et al., 2014; Ahn and Chodroff, 2022).

This dissertation thus seeks to bridge the gap between the everyday experience of individual vocal variation and the broader evolutionary dynamics of spoken language. By examining how anatomical variation influences the production and acoustic realization of speech sounds, this dissertation aims to shed light on one pathway through which biological diversity drives the ongoing diversification of languages (and of birdsong, but more on that later).

1.2 Studying speech and language: a complex problem

The study of spoken language and its origin has repeatedly been marked as "the hardest problem in science" (Christiansen and Kirby, 2003a; Christiansen and Kirby, 2003b). Language evolution is based on myriad complex processes that occur in complex social environments, including cultural, biological, and environmental mechanisms (Blasi et al., 2019; Dediu et al., 2019; Gisladottir et al., 2023; Liang et al., 2023). Additionally, a fundamental distinction often remains underexplained: while language refers to the abstract, rule-governed system of communication, speech is the physical realization of language through articulated sound. The two are inherently connected – as will be shown throughout this dissertation. Given this complexity, there is no consensus on how to methodologically approach the study of language evolution (Christiansen and Kirby, 2003b). Various approaches address different strata of language and speech to study language change – ranging from the analysis of grammar, syntax, and morphology (features of language) and how they change over time (Greenhill et al., 2017; Heggarty et al., 2023), to the investigation of acoustic-phonetic data and vocal tract morphology (features of speech) and how these affect language and language change (Moisik and Dediu, 2017; Dediu et al., 2017; Blasi et al., 2019; Dediu et al., 2019), and ultimately to the neural mechanisms underlying language, speech production, and perception (Christiansen and Chater, 2008). In this dissertation, a bottom-up approach is used to study speech, language, and their diversification – beginning with the relationship between vocal tract morphology, articulation, and acoustic output, and using this foundation to infer how the sounds of language diversify over time. This approach presents a contrast to the comparative method (Fox, 1995; Campbell and Poser, 2008; Campbell, 2013), which traditionally infers language diversification from the discrete, hierarchical units of language such as phonemes or lexical cognates across languages (Atkinson and Gray, 2005; Greenhill et al., 2017; Barbieri et al., 2022), often abstracting away from the biological and physical mechanisms of speech production. While powerful for reconstructing ancestral states and classifying language families (Gray and Atkinson, 2003; Atkinson and Gray, 2005; Heggarty et al., 2023), the comparative method typically overlooks the role of speaker-level variation and the articulatory-acoustic interface in shaping linguistic change (Dediu et al., 2017; Moisik and Dediu, 2017; Dediu et al., 2019). By contrast, the bottom-up approach presented in this work integrates morphological, articulatory, and phonetic data to ground language evolution and diversification in the physiological and cognitive realities of speech, enabling the exploration of how individual-level variability can scale up to population-level patterns of diversification (Moran et al., 2024).

However, analyzing speech itself presents a complex challenge, since it includes – if a holistic approach is warranted – the integration and analysis of the successive levels of speech production (Fig. 1.1). It is well understood how articulation, i.e., the coordinated movement of different articulators in the vocal tract to produce speech sounds, is conditioned in the brain via speech motor planning and subsequent somatosensory and auditory feedback (Bartoli et al., 2015; Gick et al., 2019; Pastore et al., 2022). The acoustic signal produced by these articulations is processed in the brain, which includes primary and secondary auditory perception in the superior temporal gyrus and successive integration of semantic and syntactic processing (Mesgarani et al., 2014; Guenther, 2016). In contrast, the connection between vocal tract anatomy and articulation, and successively between articulation and the corresponding acoustic signal is less well understood (Lammert et al., 2011). Especially

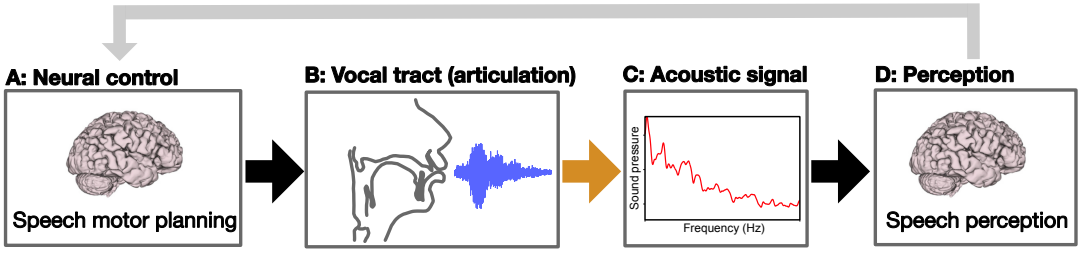


FIGURE 1.1: **The successive levels of speech production**. **A**: Speech production begins with speech motor planning in the brain, which translates into articulatory gestures of the vocal tract (**B**) (Gick et al., 2019). Vocal tract anatomy and individual articulatory gestures of speech sounds (such as vowels and consonants) produce an acoustic signal (**C**; depiction of the author's frequency spectrum while saying the word "language") (Fant, 1960), which is mainly perceived in the superior temporal gyrus of the brain (**D**) (Mesgarani et al., 2014). Finally, speech motor control is influenced by speech perception through feedback via sensorimotor integration (gray arrow) (Bartoli et al., 2015; Ekström, 2022; Pastore et al., 2022). This dissertation is mainly concerned with the processes that translate articulatory gestures and vocal tract variability into acoustic signals (**B**–**C**), and eventually in the processes that lead to language diversification, given the acoustic variation that is produced by individual articulation (orange arrow). The 3D model of the brain is from the Allen brain map.

the impact of interindividual variation on articulation and the resulting acoustic realizations is often neglected (Fuchs et al., 2008; Weirich and Fuchs, 2011; Dediu et al., 2019).

In this dissertation, the focus lies on specific aspects of speech production: the relationship between vocal tract morphology and its downstream effects on articulation and the resulting speech sounds – vowels and consonants – as expressed through their acoustic realizations (Fig. 1.1 B – C). The mapping between articulation and acoustic output has long been a subject of debate, and there remains no unified theoretical or empirical framework to resolve it (Blumstein and Stevens, 1979; Browman and Goldstein, 1992; Galantucci et al., 2006; Neiberg et al., 2008; Noiray et al., 2014; Perkell and Klatt, 2014). Proposed confounding factors include potentially non-linear relationships between articulation and acoustics (Stevens, 1972), motor equivalence in articulatory gestures (Perrier and Fuchs, 2015), i.e., different articulatory gestures may produce the same speech sound, anatomical variation across speakers (Lammert et al., 2011), the influence of biological and psychological factors like sex or gender (Simpson, 2002; Simpson, 2009), challenges in measurement and data representation (Stone, 1997; Duckworth et al., 2011), and mismatches between speech production and perception (Liberman and Mattingly, 1985). This uncertainty hampers our understanding of how morphological diversity among speakers influences spoken language and mutual intelligibility. Moreover, a clearer perspective on the articulation-to-acoustics mapping process may provide important insights into the dynamics of language acquisition, speech pathology, and ultimately, into the evolution of spoken languages. As mentioned before, further complicating this issue is the vast diversity of methodological approaches for measuring both the articulatory and acoustic signals (Fant, 1960; Wood, 1982; Stone, 1997; Fitch and Giedd, 1999; Stevens, 2000; Zollikofer et al., 2011; Gully, 2021; Yang, 2021). To address this, a further aim of this dissertation is to introduce a unified analytical framework – a toolbox for the statistical co-analysis of articulatory and acoustic variation – that

enables more systematic, comparable, and reproducible investigations across individual speakers. The details of these procedures are introduced in sections 1.4, 1.5 and 1.6.

Building on this framework, the vast variation observed in acoustic signals (Ohala, 1989; Johnson et al., 1993) – shaped by complex articulatory movements of the vocal tract – can serve as a basis for investigating the processes underlying language diversification (Fig. 1.1 C–D), and potentially the diversifying mechanisms in the vocalizations of other species. Specifically, this includes tracing how interindividual variation emerges and propagates across successive levels of speech production: from anatomical differences in the vocal tract, to articulatory variability, to the resulting phonetic patterns within speaker communities, and ultimately to large-scale phonetic variation across languages (Moran et al., 2024; this will be discussed in detail in Chapters 2, 4 and 5).

In summary, this dissertation aims at offering a fresh perspective on the integrated analysis of articulatory and acoustic data, with the overarching goal of advancing our understanding of how such data can be used to explore the dynamics of language diversification. The following Sections (1.3 – 1.8) provide the theoretical background necessary to contextualize the methods and conclusions presented in the various chapters of this dissertation.

1.3 The vocal tract

To understand the physical basis of speech and its role in language evolution, we now turn to the anatomy of the vocal tract – the biological system that underlies articulatory movements and ultimately shapes the acoustic signal. Speech is produced in the vocal tract, a system of anatomical structures and muscles that connect the lungs to the upper vocal tract, including the lips and nose (Fig. 1.2). These structures – collectively called speech organs – are anatomical adaptations and exaptations that can coordinate at millisecond precision to produce speech (Ghazanfar and Rendall, 2008; Simonyan and Horwitz, 2011; Falzone et al., 2014). During speech, exhalation is modulated by abdominal muscle tension, which expels air from the lungs through the trachea, glottis, and into the supralaryngeal vocal tract – the region above the larynx encompassing the oral and nasal cavities (Gick et al., 2013).

The larynx, a complex cartilaginous and muscular structure atop the trachea, separates the lungs from the supralaryngeal tract. Beyond its role in breathing and protecting the airway during swallowing, it functions as the motor of speech. As subglottal air pressure builds below closed vocal folds, it eventually overcomes their tension, causing the folds to part and air to flow through (Gick et al., 2013). This airflow lowers pressure across the folds (Bernoulli effect), drawing them back together and initiating oscillation – described by the myoelastic aerodynamic theory of phonation (Van den Berg, 1958). Variations in vocal fold tension, position, and airflow produce a range of sound sources – periodic, aperiodic, or mixed – enabling different phonatory modes such as modal, creaky, and breathy voicing (Gick et al., 2013). These modes contribute to the rich variability of human vocal communication.

Finally, when the airflow reaches the pharynx, the rest of the vocal tract functions as a resonator, transforming aerodynamic energy into acoustic energy. Air pressure variations generate sound waves (Fry, 1979), which – when within the human auditory range – are perceived as sound. The specific vocal tract shape during articulation – such as the vowel gesture for the vowel [u] and the gesture for the

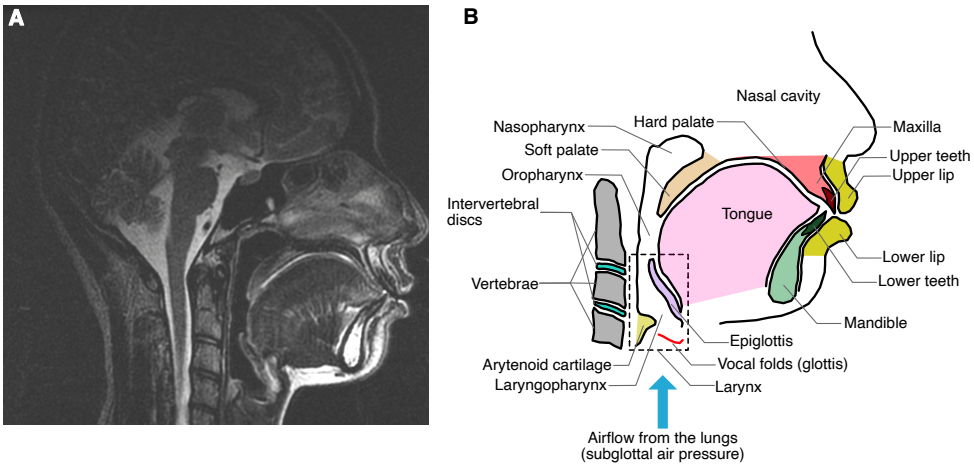


FIGURE 1.2: **Vocal tract imaging and anatomy**. **A**: Example of a static MRI (Magnetic Resonance Imaging) of the vocal tract of a female speaker in the midsagittal plane, depicting the major structures involved in articulation. **B**: Anatomical description of the major structures (same speaker), which may be traced in midsagittal MRI of the upper vocal tract. Note that in real-time MRI image data as compared to static MRI data, fine structures such as the vocal folds might not be visible. Additionally, note that only part of the larynx is visible in this MRI image (**B**; see dashed box). The MRI image data is from the USC 75-Speaker Speech MRI Database (Lim et al., 2021).

consonant [t] (see Fig. 1.3) – selectively reinforces certain frequencies and attenuates others (Fant, 1960). Its configuration, together with laryngeal settings and articulator constrictions (via the tongue, lips, teeth, palate), underlies both innate vocalizations like crying and laughing (Anikin et al., 2018), and the extensive phonetic diversity, i.e., the diversity of speech sounds found in the world's languages (Ladefoged and Maddieson, 1996; Moran and McCloy, 2019). Further details on articulation and its acoustic consequences are provided in Section 1.5.

However, while this dissertation primarily focuses on humans, we are by no means the only species with vocal organs to produce complex vocalizations. Various other mammals, birds, and fish can produce sophisticated vocal signals using complex vocalization procedures (Bell et al., 2004; Fitch, 2006; Horvatić et al., 2021). A particularly interesting analogue to humans is the vocal capability of songbirds (Doupe and Kuhl, 1999; Zhang et al., 2023). In songbirds, vocalizations are also produced by expiration, but through an organ called the syrinx. The syrinx can produce more varied and more complex sounds than the human larynx because the syrinx has two independently controllable sound sources, i.e., one to each lung (Suthers, 2004). This allows, for example, the Marsh Warbler (*Acrocephalus palustris*), a small passerine songbird, to mimic the songs of many other species, e.g., Swallows, Thrushes, Nightingale, resulting in a repertoire that far exceeds 1000 distinct syllables (Dowsett-Lemaire, 1979; Bell et al., 2004). More on the vocal capabilities and the parallels between acoustic diversification in birds and humans is presented in Chapter 4.

From an evolutionary perspective, it is important to briefly introduce the morphology of the human vocal tract in comparison to that of other great apes, as they are our closest extant relatives and provide key insights into the anatomical and functional changes that may have supported the evolution of speech and language.

A striking difference is that humans possess a descended larynx – our larynx is positioned lower in the throat than in other great apes (Fig. 1.2; cf., Fitch, 2000). However, this is a feature we also share with some non-primate mammals, such as deer (Fitch and Reby, 2001). While human infants are born with a high-positioned larynx that allows for simultaneous breathing and suckling by connecting the larynx to the nasal cavity (Crelin, 1987), the larynx gradually descends into the throat over the first few years of life as part of normal anatomical development (Fitch and Giedd, 1999). This anatomical trait was historically believed to be essential for producing a wider range of speech sounds, especially vowels (Lieberman et al., 1969). Developmental evidence supports the idea that vocal tract length and configuration have phonetic consequences: young children, who have a shorter and higher vocal tract and do not yet produce the full set of speech sounds (Fitch and Giedd, 1999; Namasivayam et al., 2020), show more limited vowel contrasts and less acoustic distinctiveness in their speech, in part due to these anatomical constraints (Chung et al., 2012; Kent and Rountrey, 2020). However, recent work disputes these findings and argues whether a descended larynx is necessary – or even advantageous – for speech production (Fitch et al., 2016; Boë et al., 2019; Janssen et al., 2019).

Alternatively, it has been proposed that the descended larynx functions primarily in the acoustic exaggeration of body size and of sex differences, as elongating the vocal tract lowers resonance frequencies and enhances the perception of largeness (Fitch, 1997; Fitch and Reby, 2001). Further, recent anatomical studies suggest that, over evolutionary timescales, the human larynx underwent structural simplification compared to that of other great apes, potentially facilitating more consistent and controllable vocalizations (Nishimura et al., 2022). However, other researchers argue that it is primarily the neural architecture supporting speech and language that underlies humans' complex vocal capacities (Fitch et al., 2016; Dubreuil and Henshilwood, 2013; Albessard-Ball and Balzeau, 2018; Balezeau et al., 2020). In summary, the literature presents conflicting views, and the question of what makes the human vocal tract truly unique remains somewhat unresolved. This issue will be revisited in Chapter 6.

1.4 Studying articulation

There is a wide variety of approaches to study articulation (Wood, 1982; Stone, 1997; Fitch and Giedd, 1999; Zollikofer et al., 2011; Gully, 2021). This dissertation presents a comprehensive methodology for measuring and analyzing articulation using MRI data. The focus is on measurement and analysis of the upper vocal tract – the active and passive articulators involved in producing vowels and consonants (Fig. 1.2).

1.4.1 Articulatory data

Various classes of imaging techniques are available for the acquisition of articulatory data, including Ultrasound (Wilson, 2014), Computed Tomography (CT) (Serrurier and Badin, 2005), X-ray microbeam (Papcun et al., 1992), Electromagnetic Articulography (EMA) (Rebernik et al., 2021), and real-time Magnetic Resonance Imaging (rtMRI) (Narayanan et al., 2014), among others (cf., Stone, 1997; Gick et al., 2013; Lin, 2021). Ultrasound is commonly used to visualize the tongue and palate during articulation, but it does not provide a holistic or high-resolution view of the entire vocal tract. EMA, on the other hand, uses electromagnetic sensors placed on articulators (e.g., tongue, lips, and palate) to track their relative movement in three dimensions

during speech. Although EMA offers feasible temporal and spatial resolution, it requires time-intensive setup procedures and frequent system calibration (Gick et al., 2013). Moreover, it is relatively invasive and has the potential to alter natural articulatory behavior (Tienkamp et al., 2024).

In contrast, both CT and X-ray microbeam imaging provide high-quality visualizations of the vocal tract. Notably, X-ray microbeam enables dynamic tracking of articulatory motion during speech tasks. However, both techniques involve exposure to ionizing radiation, though the dose in X-ray microbeam studies is relatively low (Gick et al., 2013). MRI thus presents a valuable alternative, offering high-resolution static and dynamic imaging of the entire speech apparatus without the risks associated with radiation exposure (Gick et al., 2013; Narayanan et al., 2014). Figure 1.3 illustrates an example of rtMRI during articulation of a vowel-consonant-vowel sequence [utu].

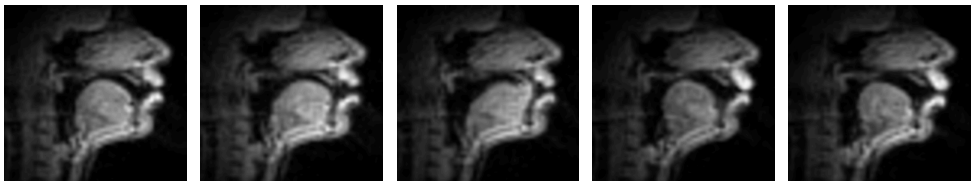


FIGURE 1.3: **Real-time vocal tract imaging**. Illustration of a dynamic series of real-time MRI images (rtMRI) in the midsagittal plane of a male speaker producing a vowel-consonant-vowel sequence (here: [utu]). For simplification only a subset (five frames) of the sequence is shown. The image data was recorded at 23.18 frames per second. The MRI image data is from the USC Speech and Vocal Tract Morphology MRI Database (Sorensen et al., 2017).

In this dissertation, rtMRI is utilized to analyze imaging data across various speech sounds, including stationary vowels and dynamic sequences such as vowel-consonant-vowel (VCV) utterances (Fig. 1.3). Real-time MRI records MRI images at rates of up to 83 frames per second (Lim et al., 2021), which permits the simultaneous fine-grained analysis of vocal tract morphology during articulation and the time-aligned acoustic realizations thereof, since it is possible to make reliable acoustic recordings directly in the MRI scanner (Bresch et al., 2006).

1.4.2 Geometric morphometrics

In this dissertation, I employ geometric morphometrics (GM) to quantitatively evaluate articulatory shape using real-time MRI (rtMRI) data collected during speech (see Fig. 1.4; Bookstein, 2018). GM is an analytical toolkit used for the quantitative analysis of shape of two- or three-dimensional biological structures (Mitteroecker and Gunz, 2009; Zollikofer et al., 2011; Zelditch et al., 2012; Klingenberg, 2015; Dryden and Mardia, 2016; Bookstein, 2018; Ponce de León et al., 2018; Alatorre Warren et al., 2019; Zollikofer et al., 2022). Traditional morphometrics measures linear distances (e.g., length, width, height) on biological objects and uses multivariate statistics to analyze variation among or within groups (Lagler et al., 1962; Strauss and Bookstein, 1982; Marcus, 1990). This approach has also been applied to articulatory data (Fitch and Giedd, 1999). However, linear distance measurements are problematic because they are often correlated with size, they often do not rely on homologous anatomical reference points, and make it difficult to graphically represent shape. Additionally, identical measurements can arise from different shapes

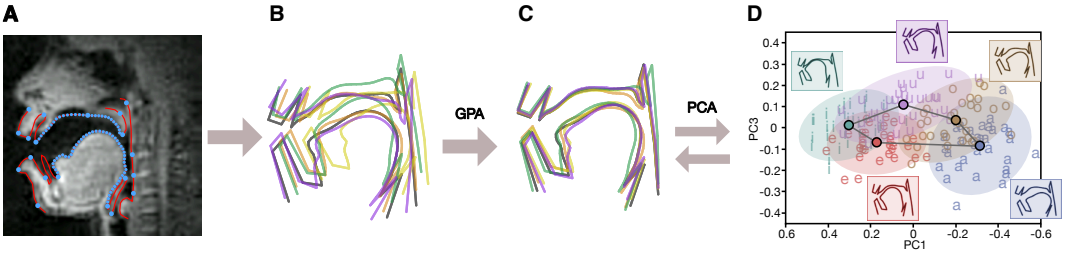


FIGURE 1.4: **Geometric morphometrics of the vocal tract**. **A**: Anatomical landmarks are placed on MRI image frames to quantify vocal tract shape (example: vowel [a]). Fixed landmarks (large blue dots) mark homologous anatomical points; semi-landmarks (small blue dots) follow curves such as the palate. Red lines trace major vocal tract outlines. **B—C**: Raw landmark configurations (**B**) vary in position, scale, and orientation. Generalized Procrustes Analysis (GPA) aligns them into a shared coordinate system (**C**), preserving only shape variation. **D**: Principal Component Analysis (PCA) of GPA-aligned shapes visualizes articulatory variation in multivariate space. A key strength of geometric morphometrics is that abstract shape differences (PCs; **D**) can be mapped back to physical landmark configurations (**C**), enabling interpretation in both shape and anatomical space.

---

if the spatial relationships between anatomical reference points are not considered (Adams et al., 2004).

GM overcomes these limitations by quantifying shape with geometric coordinates instead of linear measurements (Fig. 1.4 A). Compared to traditional multivariate morphometric methods, the major advantage of modern GM is that complex patterns of shape variation can be analyzed simultaneously in abstract multivariate space, i.e., the shape space (Fig. 1.4 D), and in the original physical space of actual morphologies (Fig. 1.4 C). Thus, GM and related multivariate methodologies can be used for the co-analysis of articulatory and acoustic data (cf., Geng and Mooshammer, 2009; Chang et al., 2014; Narayanan et al., 2014).

A geometric morphometric (GM) analysis follows a series of standardized steps known as the Procrustes paradigm (Fig. 1.4 A–C); Adams et al., 2013). This approach unifies the statistical treatment of shape (Kendall, 1981; Kendall, 1984; Kendall, 1985) by aligning landmark configurations through Generalized Procrustes Analysis (GPA; Gower, 1975; Rohlf and Slice, 1990; Dryden and Mardia, 2016). The GM workflow typically involves: (1) collecting landmark data, (2) performing GPA to remove differences in position, scale, and rotation, and (3) conducting multivariate analyses on the resulting shape variables. The next few paragraphs outline the steps of a standard GM procedure.

First, landmarks are placed on anatomically well-defined points (so-called landmarks) to sample aspects of the shape of interest (see 2.S1 in Chapter 2 and Fig. 1.4 A), e.g., points of biological significance or functional homology that can be replicated from object to object (fixed landmarks). Fixed landmarks are clearly defined points that indicate particular anatomical traits and discrete biological attributes, such as the tip of the front teeth (Fig. 1.4 A). Additionally, semi-landmarks along curves or outlines can be placed between two fixed anatomical endpoints (fixed landmarks) in order to capture shape variation of outlines and curved objects (such as the outline of the hard palate or the tongue; see Fig. 1.4 A; Mitteroecker and Gunz, 2009). Curve semi-landmarks are then equidistantly sampled and are are allowed to slide along tangents to the curves to ensure geometrical correspondence between all configurations (Bookstein, 1996; Gunz et al., 2005). From a digitized image, such as

an MRI, fixed landmarks as well as points along curves can be obtained with common software packages and libraries in *R* (Olsen and Westneat, 2015; R Core Team, 2024).

Second, the extracted landmark data typically contain non-relevant variation in terms of orientation and position as well as biologically relevant variation in size. Non-relevant variation arises from differences in positioning and head posture during scanning, or because of inconsistencies in image alignment across individuals. Hence, before any further analysis is undertaken, non-relevant variation is removed via Generalized Procrustes Analysis (GPA; Dryden and Mardia, 2016). GPA translates and rotates individual landmark configurations to minimize their squared deviations from the sample mean shape. Additionally, it scales each configuration to unit centroid size, thereby removing differences in position, orientation, and size. To clarify, in this context, shape refers to the geometric relationships among landmarks once differences in location, orientation, and size have been removed, while size is captured by a single scaling factor – centroid size – defined as the square root of the summed squared distances of landmarks from their centroid.

Third, the GPA data are used to evaluate a sample mean shape, which serves as a reference for all further analyses. Each subject's shape is expressed as its landmark-by-landmark deviation from the reference landmark configuration (Fig. 1.4 C–D; Zelditch et al., 2012). The resulting multidimensional shape data can then be analyzed with various statistical procedures, including Principal Component Analysis (PCA), Canonical Variates Analysis (CVA) and Multiple Multivariate Regression (MMR) (Mitteroecker and Gunz, 2009). PCA is the most commonly used technique and can be computed on the shape variables to obtain a low-dimensional representation of shape data (Fig. 1.4 D). PCA yields linearly uncorrelated (i.e., orthogonal) dimensions that represent the major variation in the data, the so-called Principal Components (PCs), accounting for largest, second-largest, etc. proportions of the total variation in the sample. Normally, the first few PCs together comprise a significant proportion of the total variation in the sample, and may be used to visualize the major patterns of shape variation (Zollikofer and Ponce de León, 2002). Since multivariate shape space and physical landmark space are equivalent to each other (Fig. 1.4 C–D), patterns of shape variation can be explored and visualized in both spaces simultaneously. Although these methods are very powerful, it must be kept in mind that PCA and other ordination techniques are statistical methods, such that they do not necessarily correspond to biologically meaningful factors. Thus, especially higher-order shape PCs should be treated with caution (Mitteroecker et al., 2005; Mitteroecker and Bookstein, 2007).

Note that scaling to unit centroid size during GPA removes linear size variation from the landmark data, but the resulting shape configurations still exhibit non-linear shape variation due to size differences between subjects – a phenomenon known as size allometry (Klingenberg, 2016). Thus, the effect of size allometry on the shape of the object under study must be taken into account in GM analyses and when evaluating hypotheses. This issue is further discussed in Chapters 2 and 3.

1.5 The acoustics of speech sounds

As described in Section 1.3, speech sounds are produced through the interaction of a sound source (the vocal folds) and a filter (the articulatory configuration of the vocal tract; Fant, 1960). The preceding sections outlined how articulatory gestures can be analyzed using geometric morphometrics (GM). In what follows, I turn to

the acoustic consequences of these gestures – examining how sound is generated in detail, and how the resulting acoustic patterns can be measured and analyzed.

The sound source originates in the larynx, where the periodic vibration of the vocal folds generates glottal pulses – acoustic waves that determine the fundamental frequency (F0) of the speech sound and its harmonics (Fig. 1.5 A–B; vocal folds, i.e., the "source" in blue). The fundamental frequency (F0) is proportional to the rate at which the glottal pulses occur, which then determines the perceived pitch of a voice (i.e., how high or low a sound is perceived; Gick et al., 2013). F0 can be individually modulated (producing higher or lower pitched voice) by altering the tension of the vocal folds through complex muscular control of the larynx (Wrench and Beck, 2021). The filter is the supralaryngeal vocal tract, whose configuration – shaped by articulatory gestures like tongue and lip position – modifies the source signal by enhancing certain frequencies (resonances) and attenuating others (Fig. 1.5 A & C; supralaryngeal vocal tract, i.e., the "filter" in red). This relationship between source and filter is formalized in the source-filter theory of speech production (Fant, 1960), which explains how vocal tract shape influences the acoustic output.

The resulting complex wave can be decomposed into its component sine waves – each with its own frequency, amplitude, and phase – via the Fast Fourier Transform, allowing the signal to be analyzed in the frequency domain, i.e., the frequency spectrum (Fig. 1.5 C; Ladefoged, 1996). The dominant resonant frequencies, or formants, appear as peaks in the power spectrum and are crucial for distinguishing vowel qualities (Stevens, 2000). The spectral envelope – the shape of the frequency spectrum (Fig. 1.5 C) – reflects both the formants and timbral characteristics, encoding speaker-specific information that arises from individual vocal tract anatomy and speaker-specific articulation patterns (Warren et al., 2005).

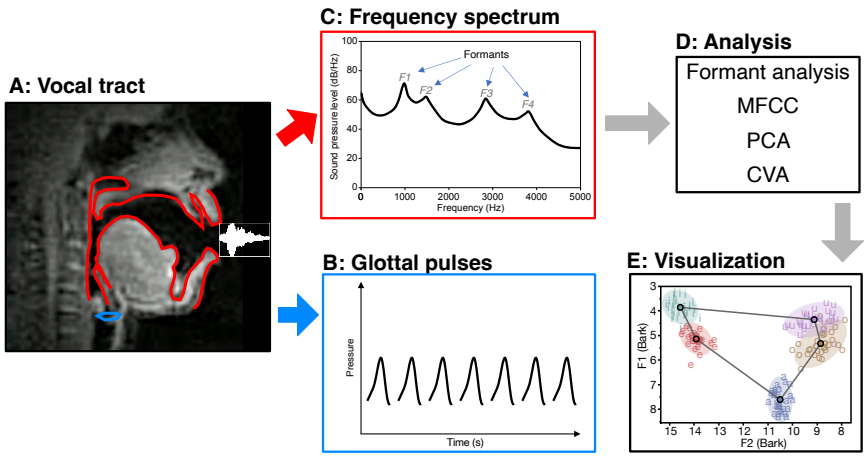


FIGURE 1.5: **Overview of acoustic phonetics**. **A**: The speech organs include the vocal folds (blue), which generate the source signal, and the articulators (red), which shape the vocal tract for specific sounds (here, vowel [a]). **B**: According to the source-filter theory (Fant, 1960), glottal pulses from the rhythmic vibration of the vocal folds create a periodic source signal. **C**: The vocal tract configuration filters this signal, producing a frequency spectrum where formants appear as spectral peaks. **D**: This spectrum can be analyzed using formant analysis, Mel Frequency Cepstral Coefficients (MFCCs; Davis and Mermelstein, 1980), or dimensionality reduction methods like PCA and CVA. **E**: Resulting acoustic variation can be visualized in low-dimensional spaces, such as the classical F1–F2 formant plot in the psychoacoustic Bark scale (Traunmüller, 1990).

Acoustic analysis techniques such as formant tracking (explained in detail in Section 1.5.1), MFCCs (explained in Section 1.5.2), and dimensionality reduction techniques, such as principal component analysis (PCA), and canonical variate analysis (CVA) (Section 1.4.2) may be used to capture and quantify this variation in a multivariate acoustic space. These methods allow the statistical analysis and visualization of speech sound differences in low-dimensional spaces, such as classical formant plots, making it possible to study both linguistic content and speaker-specific features (Fig. 1.5 E).

1.5.1 Formant analysis

The previous Section (1.5) described how articulation shapes the frequency spectrum of speech sounds, with major frequency peaks – called formants – emerging as key acoustic features (Fig. 1.5 C). These formants, especially the first two (F1 and F2), are central to identifying vowel quality and are widely used in phonetic and cross-linguistic studies (Chung et al., 2012; Yang, 2021). Formants are commonly extracted from spectrograms of speech sounds (Fig. 1.6 A; Yang, 2021), using automated formant trackers with software such as Praat (Boersma, 2001). Spectrograms visualize the dynamic nature of speech production. They plot frequency over time with an additional third dimension; the amplitude or intensity of a component frequency, which appears as dark bands, i.e., the formant frequencies (Boersma, 2014; Fig. 1.6 A; see colored lines).

Phoneticians typically visualize the extracted formants by plotting the first two formant frequencies (F1-F2) in a bivariate graph (Weckwerth, 2021; Fig. 1.6 B). Vowel quality, i.e., the distinct positions of vowels in a formant plot (Figs. 1.6 B–E) is classically described through articulatory and auditory properties (compare Fig. 1.4 D) – jaw/tongue height, frontness/backness of the tongue, and lip rounding (International Phonetic Association, 1999).

However, formant analysis often involves mapping acoustic data onto perceptual scales to better align with human auditory perception. Logarithmic scales, for example, compress frequency ranges – such that a doubling in frequency is perceived as an octave – making them more tractable and perceptually meaningful (Harrington et al., 1999). Among the most commonly used representations are the Bark and Mel scales. The Bark scale divides the audible spectrum into 24 critical bands of approximately equal perceptual width (Fig. 1.6 C; Zwicker, 1961; Traunmüller, 1990). It increases linearly at low frequencies and exponentially at higher ones, mirroring how we perceive loudness. The Mel scale, derived from pitch judgment experiments, transforms frequency logarithmically to match perceived pitch distances (Stevens et al., 1937; Stephens and Volkman, 1940). A frequency difference of 100 Hz is more salient at lower frequencies (e.g., 100–200 Hz) than at higher ones (e.g., 3000–3100 Hz), and the Mel scale accounts for this non-linear sensitivity. The Mel scale is used for the computation of Mel Frequency Cepstral Coefficients (Section 1.5.2). Taken together, these scales allow for more perceptually realistic analyses and visualizations of vowel spaces and improve feature extraction for tasks like speech synthesis, recognition, and cross-linguistic comparison (Iivonen, 1994; Hermansky et al., 2013). Depending on the research goal, formant data may be analyzed in raw frequency (Hz), Bark, or Mel scales and further explored using techniques such as PCA and CVA for dimensionality reduction.

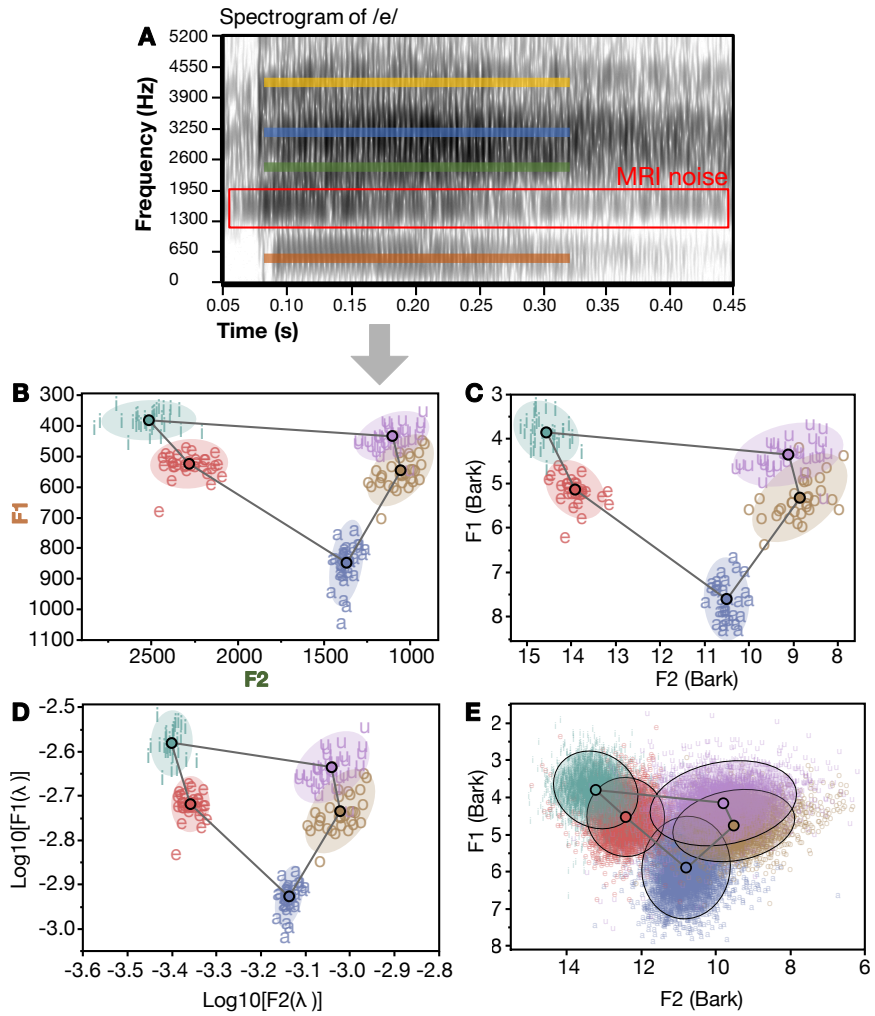


FIGURE 1.6: **Formant extraction and visualization**. **A**: Spectrogram of the vowel [e] recorded in an MRI scanner (Lim et al., 2021). Colored lines indicate formant frequencies; the strong band between F1 and F2 reflects MRI scanner noise. **B**: Classical F1–F2 plot (in Hz) for vowels i-e-a-o-u, with 90% density ellipses and mean values connected as a vowel polygon. **C**: F1–F2 values shown in Bark scale (Traunmüller, 1990), capturing the logarithmic nature of auditory perception. **D**: F1–F2 plotted as log10-transformed wavelengths (λ), as vocal tract changes are reflected more proportionally in wavelengths (Fry, 1979). **E**: Bark scaled F1–F2 plot from VoxCommunis corpus (Ahn and Chodroff, 2022), showing cross-linguistic vowel variation.

1.5.2 Beyond the peak: continuous acoustic features

In the previous Section, the classical and widely applied technique of formant analysis was introduced. However, formant analysis has various limitations, as it "only" identifies local maxima in the sound spectrum and may overlook other important articulatory and acoustic dimensions (Bladon, 1982; Holmes et al., 1997; Harrison, 2013; Maurer, 2016). Furthermore, there is an ongoing debate about the extent to which formants directly correlate with the perception of speech sounds (Bladon, 1982; Zahorian and Jagharghi, 1993; Nenadić et al., 2020). As a result, it has been proposed that listeners may rely on more global spectral features, often referred to as the "overall" or "gross" spectral shape (Kiefte et al., 2013), alongside or even instead of "discrete" formant peaks (Plomp et al., 1967; Klein et al., 1970; Zahorian and

Jagharghi, 1993; Ito et al., 2001; Friedrichs et al., 2017; Nenadić et al., 2020). In line with the multivariate and continuous nature of articulation (Fig. 1.4 D), the spectral envelope itself (Fig. 1.5 C) along with methods such as Mel Frequency Cepstral Coefficients (MFCCs) that approximate a perceptually filtered version of this broader spectral envelope, can perhaps provide a more comprehensive characterization of vowel acoustics (Davis and Mermelstein, 1980).

MFCCs are used in most state-of-the-art speech and speaker recognition systems and disease detection (Davis and Mermelstein, 1980; Tiwari, 2010; Tracey et al., 2023). MFCC analysis divides the acoustic signal into short feature vectors by analyzing short time windows. For each window, a power spectrum is computed using a Fast Fourier Transform (FFT), transforming the signal from the time domain into the frequency domain. This spectrum is then filtered through a Mel scale filter bank to reflect the non-linear frequency sensitivity of human hearing. A cepstral analysis of the Mel scaled spectrum yields MFCCs. These features mimic human auditory perception and help separate the source signal (glottal pulses) from the filter effects (vocal tract shape), with more filters concentrated in the speech-relevant frequency range to model human critical bandwidths (Logan, 2000). The classical extraction of MFCCs of a speech signal typically retains 12-13 MFCCs (Ellis, 2005), which can be visualized – similar as shape variables – in a reduced multivariate acoustic space using PCA (see Section 1.4.2). Figure 1.7A shows the five cardinal vowels in MFCC-PCA space, which closely resembles multivariate articulatory shape space (compare Fig. 1.4 D). Similarly, the raw and unfiltered frequency spectrum (Fig. 1.5 C) can be directly analyzed and visualized using PCA (Fig. 1.7 B) and CVA (Fig. 1.7 C). A comprehensive application of MFCCs is presented in Chapter 3.

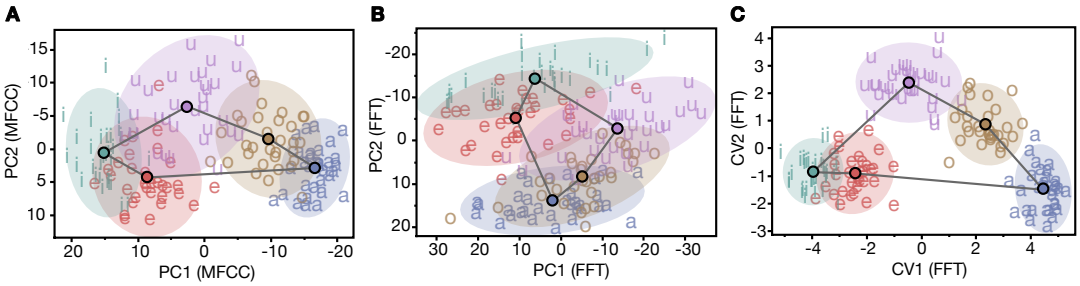


FIGURE 1.7: **Alternative acoustic measures**. **A**: Representation of acoustic variation of vowels using Mel Frequency Cepstral Coefficients (MFCCs) with subsequent visualization using PCA. The density ellipses show the 90% variation around each vowel, the colored points show the mean vowels, which are connected to represent the vowel polygon. **B**: PCA of the frequency spectrum of vowel utterances via the Fast Fourier Transform. **C**: CVA of the frequency spectrum.

1.6 Morpho-phonetics

As discussed in Sections 1.4 and 1.5, there are advanced methods for reliably measuring both the articulation and acoustics of speech sounds. However, a unified statistical approach to study the multivariate representations of these two domains and their interrelations is lacking. In this Section, I provide an overview of the co-analysis of corresponding articulatory and acoustic datasets (Fig. 1.8).

Data resources such as the 75-speaker speech MRI database from the University of California provide corresponding articulatory data (rtMRI) and audio recordings

of various speech sounds and sequences of speech sounds (Lim et al., 2021). This enables a co-analysis of the two domains of data, i.e., a simultaneous analysis of articulatory and acoustic data (Fig. 1.8 A–C). The vocal tract shape data of each individual speaker – analyzed using GM – can be directly correlated with their acoustic data. We refer to this integrated analysis of articulatory and acoustic data as *morpho-phonetics* (Moran et al., 2024), as it combines the examination of morphological vocal tract data using geometric morphometrics (GM) with the acoustic analysis of speech sounds. This methodology may be utilized to study many aspects of speech, such as effects of articulatory to acoustic correspondence (see Chapter 2), coarticulation of vowel-consonant-vowel (VCV) sequences (see Chapter 3), and potentially provides a new method for the study of speech pathology and in forensic speech science.

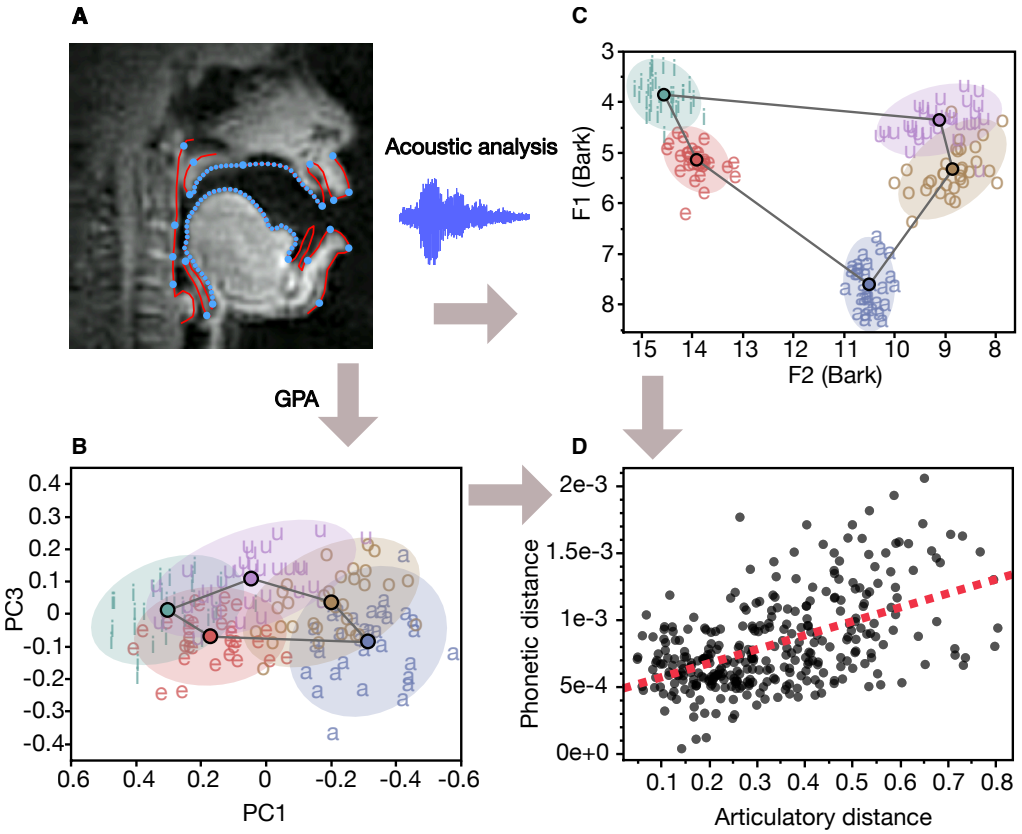


FIGURE 1.8: **Morpho-phonetics of vowel production**. **A**: Real-time MRI frame of the vocal tract of a speaker in the midsagittal plane producing the vowel [a] with corresponding anatomical landmarks. At each MRI frame and corresponding acoustic timestamp, analysis of shape via GPA and acoustic phonetic analysis is undertaken. The MRI image data is from the USC 75-Speaker Speech MRI Database (Lim et al., 2021). **B**: Shape variation of vowels in low-dimensional multivariate shape space (PCA). **C**: Acoustic variation of vowels in Bark scaled F1-F2 space (Traunmüller, 1990). **D**: The resulting corresponding shape (**B**) and acoustic (**C**) variation may be used to study the correspondence between articulatory and acoustic data. The figure shows the linear regression of the relationship between phonetic and articulatory distances among vowels.

Morpho-phonetics requires the exploration of correlation and covariation between the two time-aligned spaces (articulatory and acoustic; Fig. 1.8 B–C) and includes several multivariate statistical approaches. One advantage of GM is that the multivariate shape space of articulatory gestures (Fig. 1.8 B) may be considered analogous

to the frequency space derived from the acoustic analysis of speech (Fig. 1.8 C). The shape space of GM, the Fourier space of formant analysis and/or acoustic representations of the spectral envelope (such as MFCCs), can be represented in multivariate spaces and are thus suited for exploring the link between articulatory speech gestures and their corresponding acoustic output (Figs. 1.4, 1.6, 1.7).

One way to approach the visualization and analysis of the interrelation between articulatory and acoustic spaces is through the regression of between-vowel Euclidean distance in acoustic and articulatory spaces (Fig. 1.8 D; Moran et al., 2024). Furthermore, the covariation between the acoustic realizations of speech and vocal tract shape can be evaluated directly using multivariate association methods. The most commonly used methods for exploring patterns of covariation between multiple sets of variables are Multiple Multivariate Regression (MMR) and Partial Least Squares (PLS), which can both process the multivariate data provided by shape analysis (GM) and acoustic measurements (Mitteroecker and Gunz, 2009). However, one advantage of PLS over MMR is that PLS does not necessarily require the input variables to be uncorrelated to each other, whereas MMR has difficulty dealing with correlated predictor variables. PLS can thus be used to assess covariation between the two blocks of data. PLS is based on two-block partial least squares and uses singular value decomposition to extract pairs of components (singular warps; Rohlf and Corti, 2000; Bookstein et al., 2003; Mitteroecker and Bookstein, 2007) from the covariance matrix of the two blocks. Ultimately, PLS permits the analysis of the direct impact of shape on acoustic measurements, and – similar to PCA – to render the associated patterns of covariation directly in physical shape and frequency spaces (compare Fig. 1.4 C–D). Two applications of *morpho-phonetics* are shown in chapters 2 and 3.

1.7 Studying language diversification with articulatory and acoustic data

This introduction has provided an overview of how articulatory and acoustic signals of speech, and subsequently their interrelation, can be analyzed using a new framework that I call *morpho-phonetics*. However, the question remains: how does inter-speaker variation in the anatomy of the speech organs influence the production of speech sounds and, consequently, influence the structure and diversity of spoken languages? We can expand this question further to consider how current acoustic variation across languages might offer insights into the processes of diversification that have led to the vast array of languages we observe today. To address these questions, we must explore how morphological variation affects the production and acoustic properties of speech sounds and how this variation propagates from individuals to speaker populations, ultimately influencing language diversity (Moran et al., 2024). The final part of this introduction will focus on proposing how biological diversity may drive the ongoing diversification of languages.

Languages evolve and diversify over time, influenced by both neutral and adaptive processes (Lynch and Baker, 1993; Pawlowitsch et al., 2011; Potvin and Clegg, 2015; Bentz et al., 2018). It is known that vocal tract variation influences the acoustic output of speech (Dediu et al., 2019). The morphology of an individual's vocal tract is shaped by a combination of evolutionary processes – such as natural selection and genetic drift – as well as environmental, cultural, and social influences, including lifestyle (Christiansen and Kirby, 2003b; Puts et al., 2006; Lieberman, 2007; Ghazanfar and Rendall, 2008; Fitch, 2010; Hurford, 2011; Vorperian et al., 2011; Blasi et al.,

2019; Dediu et al., 2022; Ladoukakis et al., 2022; Nishimura et al., 2022). These factors also leave distinct acoustic traces in the voice, reflecting characteristics including speakers' age, gender, sex, and body size (Simpson, 2009; Pisanski et al., 2014; Rojas et al., 2020). Additionally, these influences shape unique speaker-specific voice qualities (Laver, 1980; Kreiman and Sidtis, 2011), making each individual's voice distinct (Cavalcanti et al., 2021). Vocal tract influences may even affect language change directly, in that it inflects biases on the usage of speech sounds due to articulatory constraints (Moisik and Dediu, 2017; Blasi et al., 2019), or in that it directly affects sound change (Dediu and Moisik, 2019).

Yet despite the significance of such individual-level variation, its potential impact on language evolution has been largely absent in studies of linguistic diversity. Even though there are notable parallels between biological and linguistic evolution (Pagel, 2017; Ladoukakis et al., 2022), most studies in language diversity overlook the role of individual and community-level variation in language change (Fox, 1995; Campbell and Poser, 2008; Campbell, 2013). Diachronic linguistics usually relies on discrete language features, like the phonemic inventories of languages or lexical cognates, to reconstruct the origins of language and language family phylogenies (Atkinson and Gray, 2005; Atkinson, 2011; Creanza et al., 2015; Barbieri et al., 2022). Additionally, phonetic studies often treat acoustic differences between individuals, influenced by their vastly variable vocal tracts, as so-called "noise" in the speech signal instead of an evolutionary signal (Lobanov, 1971; Disner, 1980; Burridge, 2023). However, from an evolutionary perspective, this variation may eventually represent the substrate for between-language diversification (Moran et al., 2024).

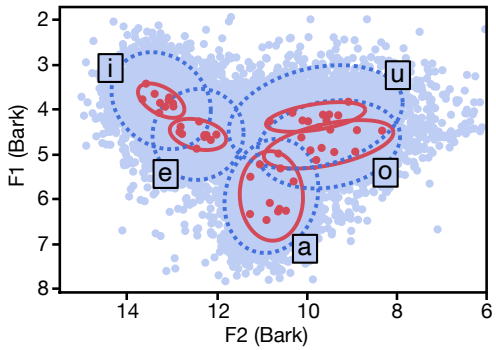


FIGURE 1.9: **Acoustic variation within- and between groups**. Representation of F1-F2 in Bark scale of a cross-linguistic sample (11 languages) from the VoxCommunis corpus (Ahn and Chodroff, 2022). The blue density ellipses show the 90% acoustic variation per vowel for the pooled data (pooled across the 11 languages), the red ellipses show the between-language 90% density ellipses. Comparing the patterns of within- vs. between-group variation can be used to gain insights into acoustic diversification of languages, such as disentangling the roles of neutral processes (e.g., drift) and adaptive pressures (e.g., cultural effects or ecological factors) in shaping language-specific phonetic features.

Given this lacuna, understanding and studying intra- and inter-group variation is crucial for studying evolutionary dynamics (Darwin, 1859; Lande, 1980; Lande and Arnold, 1983; Lande, 1991; Ackermann and Cheverud, 2004; Ponce de León et al., 2018), as it is suggested to impact sound change and divergence across groups and languages (Ohala, 1989; Mesoudi et al., 2006; MacKenzie, 2019; Anthonissen, 2021; Yu, 2021). Evolutionary theory asserts that diversification between groups feeds on variation within groups, making it crucial to understand trait variation

for explaining change (Lande, 1979). Thus, comparing variation within and across groups (speech communities, i.e., languages) may provide important insights into how languages diversify acoustically (Lande, 1980; Ackermann and Cheverud, 2004; Lande, 1991).

However, how does one go about studying diversification processes based on articulatory and acoustic variation? Firstly, by using *morpho-phonetics*, it can be inferred how overall vocal tract and articulatory variability impacts and conditions the acoustic variability within a group (see Chapter 2; Moran et al., 2024). Secondly, we can explore how the acoustic variation within groups conditions the acoustic diversification between groups by comparing acoustic variation within and between groups (compare blue vs. red density ellipses in Fig. 1.9). How articulatory and acoustic data may be utilized in detail to infer acoustic diversification of languages and of vocalization in other species such as birds is elaborated in Chapters 2, 4 and 5.

1.8 Dissertation overview

After introducing the fascinating but complex world of articulation, acoustics, and *morpho-phonetics*, this final Section provides an overview of this dissertation. The subsequent chapters present specific pieces of work that contribute to this dissertation and are grounded in the concepts and methods introduced so far. Figure 1.10 visually summarizes the structure of this work and an overall conclusion is presented in Chapter 6.

Chapter 2 serves as the foundational work of this dissertation. It explores the analysis of articulatory and acoustic data using geometric morphometrics alongside acoustic analysis techniques. It introduces the concept of *morpho-phonetics* to examine the relationship between articulatory shape space and acoustic space. Additionally, it investigates how the significant inter-speaker variation in articulation and acoustics, as well as the transformation process between articulation and acoustics, influences the ongoing acoustic diversification of spoken language by incorporating methods from evolutionary population biology (Moran et al., 2024). This chapter has been deposited as a preprint on bioRxiv.

Chapter 3 applies the methods established in Chapter 2 to test *morpho-phonetic* effects in coarticulation, which refers to the overlapping of speech sounds during fluent speech. This Chapter offers a fresh perspective on how coarticulation affects the articulation and acoustics of speech sounds as well as the acoustic diversity within languages.

Chapters 4 and 5 investigate acoustic diversity in human speech communities and in songbird populations (chaffinches), focusing on how geographic proximity and isolation (i.e., isolation by distance) shape patterns of acoustic diversification. These analyses apply methods from evolutionary population biology to both linguistic and avian data. Additionally, Chapter 5 compares different strata of linguistic structure, examining how acoustic diversification aligns – or diverges – from patterns of lexical diversification. This comparison sheds light on the extent to which acoustic divergence can reveal deeper historical relationships between languages and language families. Finally, Chapter 6 situates these findings within a broader interdisciplinary context, drawing together insights from linguistics and biology to outline implications for future research on the evolution and diversification of speech and language.

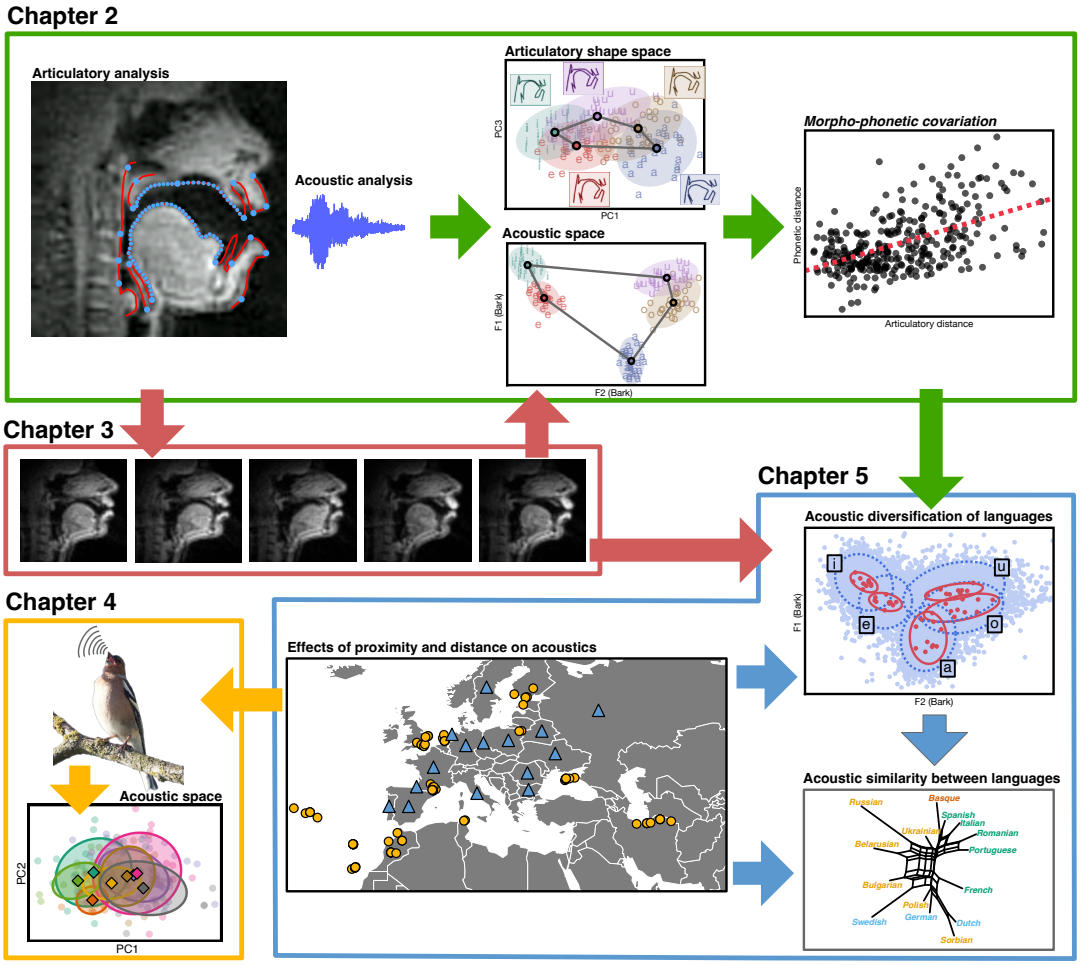


FIGURE 1.10: **Summary of this dissertation**. The MRI images are from Sorensen et al. (2017) and Lim et al. (2021). The remaining graphs and figures were created by the author. Chaffinch photo credit: author (TD).

Interindividual vocal tract diversity influences the phonetic diversification of spoken languages

2.1 Context

The following publication with the title *Interindividual vocal tract diversity influences the phonetic diversification of spoken languages* is currently in preparation for submission to a scientific journal. The preprint to this publication is available here. This study was conducted by the author in collaboration with Steven Moran (shared first authorship), Marcia Ponce de León and Christoph Zollikofer (both co-supervising authors).

2.2 Abstract

Vocal tracts, like other physical traits, exhibit extensive interindividual variation. However, little is known about how this variation translates into phonetic diversity between speakers, and ultimately, across languages. We demonstrate that different vocal tract shapes and associated articulatory strategies leave consistent acoustic signatures, which show congruent patterns of phonetic variation both within and across speech communities. Recalling a central tenet of evolutionary biology – that within-group variation feeds processes of between-group diversification – our findings suggest that the sounds of language evolve according to a neutral-like evolutionary process. Our "neutral-like evolution model" serves as a null hypothesis for disentangling the biological versus cultural mechanisms at play in phonetic diversification, where deviations from the neutral expectation indicate culturally mediated processes of language change at work.

2.3 Introduction

Textbook depictions imply a standard human vocal tract morphology (Lieberman and Blumstein, 1988; Zsiga, 2013; Ladefoged and Johnson, 2014). However, there is extensive interindividual variation in terms of the size and shape of the hard and soft tissues involved in speech production (Johnson et al., 1993; Lammert et al., 2011; Lammert et al., 2013; Wrench and Beck, 2021). Both "nature" (natural selection, genetic drift) (Puts et al., 2006; Lieberman, 2007; Ghazanfar and Rendall, 2008; Vorperian et al., 2011; Ladoukakis et al., 2022; Nishimura et al., 2022) and "nurture" (environment, culture, lifestyle) (Christiansen and Kirby, 2003b; Fitch, 2010; Hurford, 2011; Blasi et al., 2019) shape an individual's vocal tract morphology and imbue their voice with acoustic markers (age, gender, sex, size) (Pisanski et al., 2014; Rojas et al., 2020) and speaker-specific voice qualities (Laver, 1980; Kreiman and Sidtis, 2011), making them sound unique (Cavalcanti et al., 2021).

Despite considerable interindividual variation, however, speakers within a speech community reach acoustically similar targets, which they perceive as the same contrastive speech sounds, i.e., phonemes. While speaker communities thus tend to conserve phonetic (as well as other linguistic) properties, it is well known that language properties are constantly changing through time (Brochhagen et al., 2023; Greenhill, 2023), and a key question to answer is how biology, environment, and culture together bring about these changes.

Here we focus on the biological side of the question, asking specifically how vocal tract variation within a speaker community influences phonetic variation, and how phonetic variation within speaker communities gives rise to variation between communities, and eventually between languages. To the best of our knowledge, these interlinked questions have not yet been addressed comprehensively and empirically. While many studies treat acoustic differences between individuals as noise in the speech signal (Lobanov, 1971; Disner, 1980; Burridge, 2023), we are specifically interested in the patterns of variation produced by those differences. Previous studies have focused on population-level differences in vocal tract shape (Blasi et al., 2019; Dediu et al., 2019; Cavalcanti et al., 2021), on specific areas of the vocal tract (Lammert et al., 2013; Dediu et al., 2019), differences related to sex, age and size (Fitch and Giedd, 1999; Simpson, 2009; Rojas et al., 2020), have used modeling and simulation-based approaches (De Boer and Tecumseh Fitch, 2010; Lammert et al., 2011; Dediu et al., 2019; Gaines et al., 2021; Serrurier and Neuschaefer-Rube, 2023), or are preliminary (Fuchs et al., 2008; Zollikofer et al., 2011; Gully, 2021; Johnson, 2023).

In this study, we quantify the origins and propagation of interindividual variation across the successive levels of speech production; from variation in vocal tract shape and articulatory gestures to variation in the associated phonetic output within a speaker community, and from community-level to language-level phonetic variation. We use the five cardinal vowels (i, e, a, o, u) as a test case, because they are well characterized in terms of articulation and acoustics, and are shared by a wide variety of typologically diverse languages.

To quantify articulatory-acoustic relationships, we propose a new approach, *morpho-phonetics*, that combines geometric morphometric analysis (Bookstein, 2018) of the shape of the vocal tract in real-time Magnetic Resonance Imaging (rtMRI) (Lim et al., 2021) with spectral analysis of the concurrent audio signal during speech articulation (Fig. 2.1, Supplementary Fig. 2.S1). We examined the morpho-phonetic covariation of the cardinal vowels as pronounced by 32 speakers (16 women and 16 men) of American English from different ethnic and linguistic backgrounds. Vowels were extracted from the carrier words ('beat', 'bait', 'pot', 'boat', 'boot') from the USC 75-Speaker Speech MRI Database (Lim et al., 2021).

For each vowel spoken by each individual, the articulatory vocal tract shape and its corresponding phonetic output were sampled at synchronous time stamps. Physically, vowel-specific articulation is achieved by coordinated jaw, lip, tongue and velum movements, whose interplay determines the shape of the vocal tract and its acoustic output. Here, vocal tract shape is quantified with a set of anatomical landmarks (N=95) (Fig. 2.1A, Supplementary Fig. 2.S2, Supplementary Table 2.S1) placed on each rtMRI frame, and shape variation is analyzed and visualized with methods of geometric morphometrics (Fig. 2.1, A, D, E). Articulation results in an acoustic signal (Fig. 2.1, A, B, C), which is quantified here with vocal tract resonance frequencies, i.e., formants F1-F4, and the fundamental frequency, F0.

Geometric-morphometric analysis of the vocal tract as well as formant analysis yield multidimensional data sets. We use principal components analysis (PCA) as a

dimension reduction technique to explore the data structures and to characterize the principal patterns of variation in the data. Covariation between articulatory gestures and phonetic output (i.e., morpho-phonetic covariation) is examined with partial least-squares analysis (PLS). We focus on those patterns of variation that are shared by all individuals, independent of sex, age, and vocal tract size. The workflow and the main results of these analyses are visualized in Fig. 2.1; details are provided in Materials and Methods.

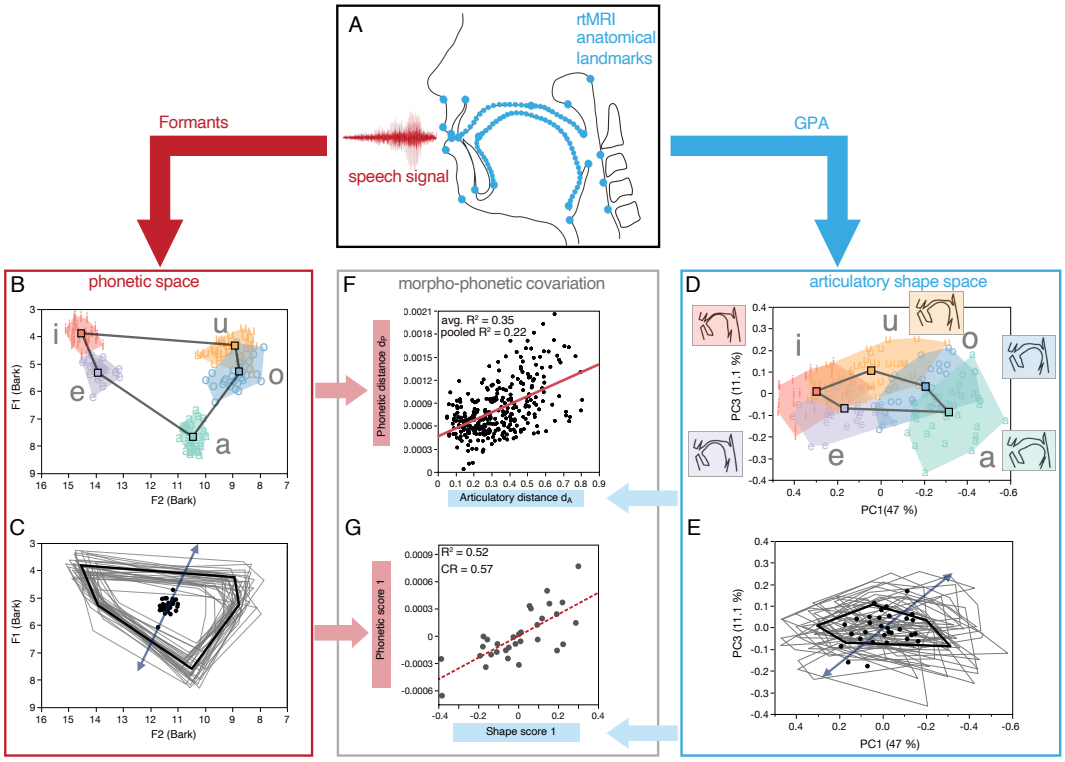


FIGURE 2.1: **Morpho-phonetic co-analysis of vowel production**. **A**: Anatomical landmarks (blue dots) are placed on midsagittal MRI images of the vocal tract recorded during speech production. At each corresponding timestamp, acoustic analysis (formant frequency extraction) is undertaken on denoised audio recordings. **B**-**C**: Resulting acoustic phonetic variation in formant space rendered as interindividual variation of vowels (**B**) and vowel polygons (**C**). Centroids of individual acoustic vowel polygons (black dots; **C**) serve as proxies of acoustic timbre. **D**-**E**: Corresponding variation in articulatory shape space (given by principal components PC1 and PC3) rendered as interindividual variation of vowels (**D**; inset graphs represent mean shape of vocal tract for each vowel) and vowel polygons (**E**). Centroids of individual articulatory vowel polygons (black dots) serve as proxies of the resting-state vocal tract morphologies. In **D** and **B**, note the reduction of within-vowel variation relative to between-vowel variation in phonetic space compared to articulatory shape space, which reflects articulatory adaptation to target vowels. **F**-**G**: Morpho-phonetic covariation. **F**: Linear regressions of phonetic vs. articulatory distances between vowels (individual regressions with average $R^2 = 0.35$; pooled regression with $R^2 = 0.27$ (red line)). **G**: Partial least-squares (PLS) analysis of the individuals' phonetic vs. articulatory vowel polygon data, reflecting residual morpho-phonetic covariation not compensated by articulatory adaptation. Details in main text and Materials and Methods.

2.4 Results

2.4.1 Articulatory and acoustic vowel spaces are similar

Figure 2.1B shows the acoustic variation given by the first two formants (F1 and F2) in Bark scale (Traunmüller, 1990). The five vowels are represented by separated data clusters and their centroids form together the well-known acoustic vowel polygon. Figure 2.1D shows the associated articulatory variation along principal components PC1 and PC3 of shape space (58.1% of total variation). PC2 accounts for interindividual differences unrelated to vowel articulation (13.3% of total variation; see Supplementary Fig. 2.S3 A). The articulatory shape space shows partially overlapping vowel data clusters, but notably the centroids of the clusters also form a vowel polygon, which has a similar geometric arrangement as the acoustic vowel polygon. The articulatory and acoustic vowel polygons therefore have corresponding shapes. To quantify the correspondence between articulatory and acoustic vowel polygons, we study how an individual's articulatory discrimination between vowels translates into acoustic discrimination. To do so, we compute for each individual all ten between-vowel distances in both articulatory and acoustic spaces. Figure 2.1F shows that articulatory distances correlate with acoustic distances (average of individual speaker correlations: $R^2 = 0.35$; pooled data: $R^2 = 0.22$).

2.4.2 Articulatory gestures are adapted to speaker-specific vocal tract morphology

The moderate correlation between articulatory and acoustic contrasts (Fig. 2.1 F) reflects the fact that the individuals in our sample have widely different vocal tract morphologies, and they use a remarkably wide range of articulatory gestures to produce any given target vowel (see Supplementary Figs. 2.S4–2.S8). Therefore, we investigate how individual vocal tract morphology influences speakers' articulatory gestures, and how these gestures translate into acoustics. A comparison of Figures 2.1B and 2.1D shows that there is substantially more articulatory than acoustic variation around corresponding vowel centroids. To quantify this effect, we evaluate the ratio of within-vowel to between-vowel variance (variance ratio; VR). We find that VR is consistently larger in articulatory space ($VR_{articulatory} = 0.548$) than in acoustic space ($VR_{formant} = 0.065$; see also Supplementary Table 2.S2). Thus, a wide range of individual articulatory gestures results in a more narrow range of acoustic variation, supporting earlier evidence that individuals adapt their articulatory gestures to produce similar sounding vowels (Lindblom and Sundberg, 1971; Ladefoged et al., 1972). To further explore the inner workings of articulatory adaptation, we examine how an individual's palate shape influences the shape of their tongue during vowel production (Fig. 2.2). To do so, we calculate the mean shape of each individual's palate and tongue, averaged over all vowels. PLS analysis shows a significant influence of palate shape on tongue shape (Fig. 2.2A, see Methods). Interindividual differences in palate shape therefore translate into systematic differences in articulatory tongue shape. Specifically, individuals with short, rounded hard palates produce more rounded tongue gestures than individuals with elongated and shallow palates (Fig. 2.2B). Fig. 2.2C indicates that articulatory adaptation of the tongue tends to maintain a standard cross-sectional air tube profile in the anterior region of the oral cavity, but less so in the posterior region. Overall, differences in vocal tract morphology are only in part acoustically compensated through articulatory adaptation.

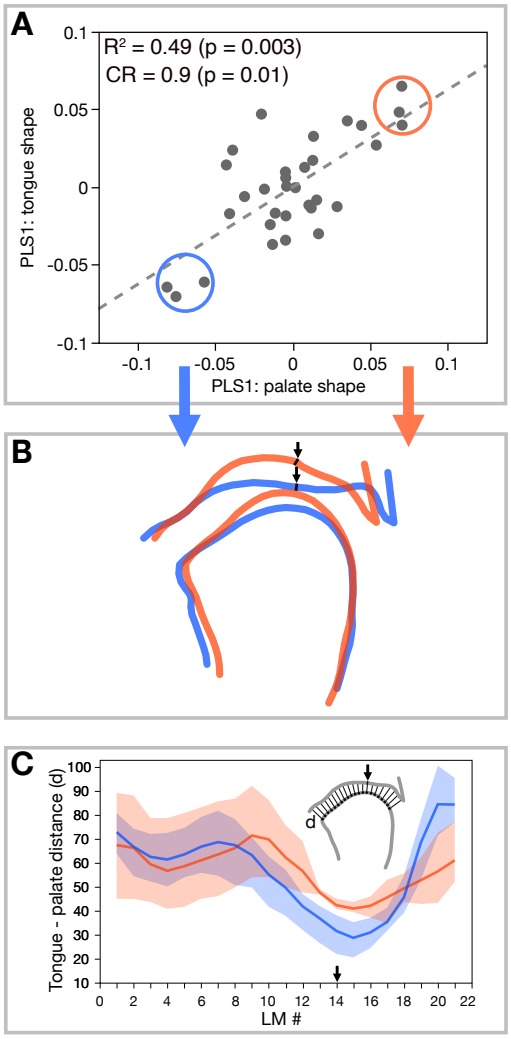


FIGURE 2.2: **Articulatory adaptation of the tongue to individual palate shape**. **A**: Covariation between tongue and palate shapes is computed with partial least squares analysis (PLS). The first pair of PLS components (PLS1) comprises 67.5% of the total covariation. **B**: Articulatory adaptation is visualized for the opposite ends of the morphological variation in **A** (average morphologies of the three individuals comprised in red and blue circles, respectively); black arrows indicate the border between the hard and soft palate. Note that in a short/rounded palate (red) the dorsum of the tongue assumes a more rounded shape. **C**: Air tube cross-sectional profiles derived from the morphologies in **B** (red/blue lines and bands represent group-specific averages and ranges). LM # indicates landmark position along tongue dorsum; **d** denotes tongue-palate distance shown in the inset graph.

2.4.3 Individual vocal tract morphology results in acoustic differences between speakers

Next we ask how the residual, uncompensated acoustic differences between individuals are related to the underlying differences in their vocal tract morphologies. To do so, we investigate the covariation between individual vowel polygon geometries in articulatory and formant space (Fig. 2.1G). A lack of covariation would indicate that the acoustic properties of the target vowels are independent of their individual articulatory properties, thus fully compensated for individual vocal tract differences. However, PLS analysis shows significant covariation (Fig. 2.3A, see Methods), confirming that articulatory adaptation is only partial (Fig. 2.2B, C), such that individual differences in the vocal tract (Fig. 2.3A) lead to systematic acoustic differences between speakers (Fig. 2.3B). This effect is visualized by the real-life morphological and acoustic properties of individuals at opposing ends of the data scatter in Figures 2.3B and 2.3C: those with rounded/short versus shallow/elongated palates. Specifically, differences in palate shape result in different timbre of the voice (as indicated by shifts in the position of the vowel polygon in acoustic space), different acoustic range (change in the size of the polygon), and different acoustic contrasts between vowels (change in the shape of the vowel polygon; Fig. 2.3, B-C). In sum, our analyses show that interindividual variation in vocal tract morphology generates correlated variation in vowel acoustics.

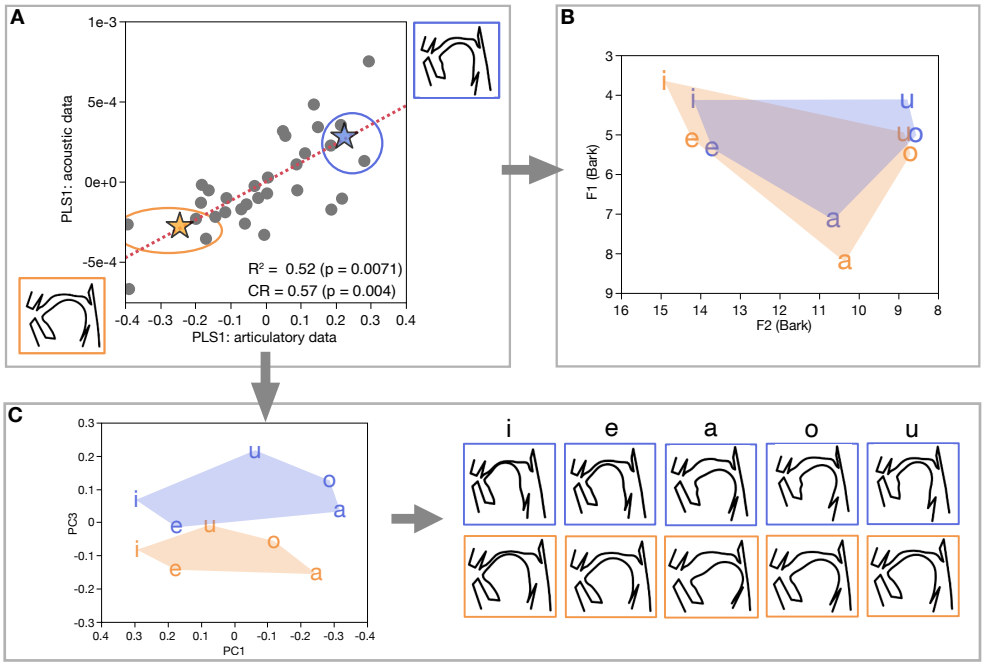


FIGURE 2.3: **Patterns of uncompensated morpho-phonetic covariation. A**: PLS analysis of the covariation between articulatory and acoustic vowel polygons (same data as in Fig. 2.1G). The first pair of PLS components (PLS1) comprises 49.8% of the total covariation. Each data point represents joint articulatory and acoustic data of one individual. Orange/blue stars represent the mean of individuals at the opposite ends of the data scatter (individuals in orange/blue circles) and are used for the visualization of contrasting articulatory morphologies (**C**) and corresponding acoustics (**B**). **B**: Corresponding acoustic vowel polygons in F1-F2 formant space (Bark scale). **C**: Articulatory vowel polygons and associated articulatory differences between individuals with rounded/short and shallow/elongate palates.

2.4.4 Vocal tract variation influences within- and between-language phonetic diversity

Figure 2.3 has shown for our morpho-phonetic sample of English speakers (MP-E) that vocal tract variation systematically conditions vowel variation. Yet the more general question remains: how does vocal tract variation and the ensuing acoustic variation between speakers ultimately influence phonetic diversification across languages? Given the current lack of cross-linguistic morpho-phonetic data, we address this question by combining the MP-E data with the multilingual speech data in the VoxCommunis Corpus (VC-C) (Ahn and Chodroff, 2022). We analyze phonetic variation of the cardinal vowels in ten typologically diverse languages of the VC-C sample (Basque, Bulgarian, Hindi, Italian, Portuguese, Romanian, Russian, Tamil, Turkish, Vietnamese) and compare it with the corresponding variation in the MP-E sample. First, we pool the data of the 10 VC-C languages to evaluate within-language variation in the sample (see Methods). Figure 2.4 shows that the pattern of within-language variation in the VC-C sample (grey ellipses) is similar to that in the MP-E sample (dashed grey ellipses) (multivariate comparison of variances: p=0.64; see Methods). Since the MP-E phonetic variation is conditioned by vocal tract variation (Fig. 2.3A), we infer that there exists a common mode of phonetic variation within languages that ultimately reflects a common mode of vocal tract variation. Second, we use the centroids from the 10 languages to evaluate the pattern of between-language variation in the VC-C sample (Fig. 2.4, red ellipses). Together, these analyses show that variation between languages is similar to variation within languages (red and grey ellipses in Fig. 2.4, respectively) (multivariate comparison of variances: p=0.36), suggesting that similar processes govern the phonetic diversification within and between languages.

2.5 Discussion

Many studies have taken a top-down approach to investigate language diversification, using methods of phylogenetic inference to reconstruct language family trees (Gray and Jordan, 2000; Holden, 2002; Gray and Atkinson, 2003). As a complement, we take a bottom-up approach to track how individual vocal tract variation in a speaker group serves as a source of articulatory and phonetic variation, and how phonetic variation within groups serves as a source of phonetic diversification between groups.

At the individual level, we found that articulatory gestures and corresponding acoustics have similar vowel polygon representations in morphometric and phonetic spaces, respectively. While articulatory vowel polygons have not yet been reported, broadly similar vowel polygon representations have been found in neuromotor and neurosensory analyses (Chang et al., 2010; Scharinger et al., 2011; Bouchard et al., 2013a; Mesgarani et al., 2014; Bouchard et al., 2016; Cheung et al., 2016; Chartier et al., 2018; Preisig et al., 2022; Oganian et al., 2022). Further studies are needed to elucidate the complete sequence of vowel representations, from neuromotor pattern generation to articulatory movements, acoustic output and speech perception (Leonard et al., 2023; Khanna et al., 2024).

We have also shown that the morphology of an individual's vocal tract leaves a consistent signature on their articulatory gestures and the associated phonetic output. Previous studies have shown that speech sounds convey information about the physical, psychological, and social characteristics of the speaker (Ladefoged and

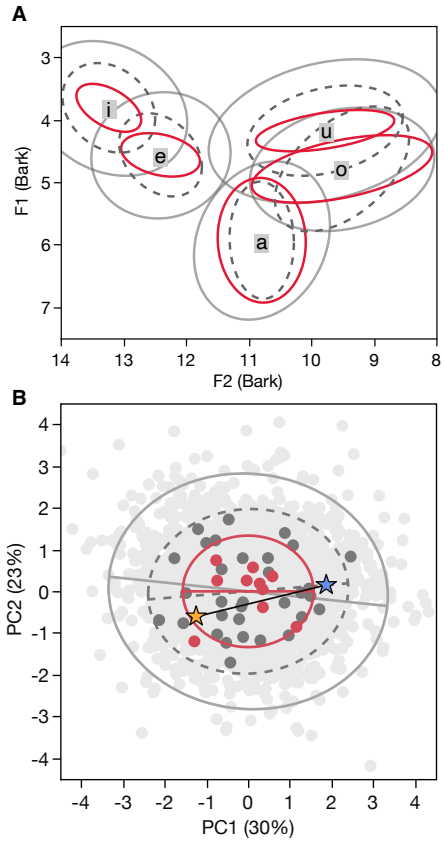


FIGURE 2.4: **Patterns of within-language and between-language variation in cross-linguistic speech acoustics data. A**: Acoustic variation of the cardinal vowels in F1-F2 formant space. 90% density ellipses show variation for the morpho-phonetic English speaker data (MP-E) (dark grey dashed lines), the pooled within-language data from the VoxCommunis Corpus (VC-C) (light grey solid lines) and the between-language VC-C data (red). Note the similarity in direction of variation across the different groups, which indicates a common mode of variation across the different data subsets. **B**: PCA of vowel polygon variation (90% density ellipses with major axes; colors as in **A**). Grey dots represent individuals (light grey: VC-C, dark grey: MP-E); red dots represent VC-C language means. The line connecting the yellow and blue stars marks phonetic variation due to vocal tract variation (MP-E data from Fig. 2.3A).

Broadbent, 1957; Kreiman and Sidtis, 2011). This morpho-phonetic signature studied here is statistically independent of sex and vocal tract size (see Methods) and represents a pattern of inter-speaker variability common to all speaker groups studied here (Fig. 2.4). Precisely because of its ubiquity, inter-speaker phonetic variability is typically regarded as "noise" that hampers categorization tasks (Johnson and Sjerps, 2021) (e.g., automatic speech recognition) as well as phyletic reconstruction (Bouchard-Côté et al., 2013). We argue that this so-called noise is an important source of phonetic variation between speaker communities, recalling a central tenet of evolutionary biology: within-group variation feeds processes of between-group diversification. Specifically, populations diversifying under a neutral (random) evolutionary process are expected to exhibit similar patterns of within- and between-population variation (Lande, 1980; Ackermann and Cheverud, 2004). In our case, variation in vowel phonetics within ($V_{\text{pho.w}}$) and between ($V_{\text{pho.b}}$) languages are indeed proportional to each other (Fig. 2.4A and Supplementary Fig. 2.S9; see Materials and Methods). Furthermore, we showed that within-group phonetic variation

is proportional to vocal tract variation (Fig. 2.3). Together, our results support the hypothesis that the morpho-phonetic covariation revealed in Fig. 2.3 is a universal biological constraint that not only shapes the phonetic diversity within speaker communities, but also governs a process of neutral-like phonetic diversification between languages (Fig. 2.4).

The neutral-like vowel evolution (NVE) model that we propose here is meant to serve as a null hypothesis that helps to disentangle the biological versus cultural mechanisms governing the evolution of spoken languages (Bentz et al., 2018). While there is consensus that phonetic and phonological change within and diversification among languages is largely driven by socio-cultural factors (Labov, 2001), empirical evidence has led to the hypothesis that changes in vocal tract morphology (be it through genetic, environmental and/or cultural factors) can change the phonological repertoires of languages (Haudricourt, 1961; Dediu and Ladd, 2007; Everett et al., 2015; Blasi et al., 2019). Model considerations propose that small vocal tract differences between speaker groups result in small phonetic differences (so-called "weak biases"), which tend to become amplified through transgenerational learning (Dediu et al., 2019). Weak-bias amplification (WBA) models typically assume that all individuals of a speaker group have the same vocal tract morphology (Dediu et al., 2019). However, our empirical data do not support this assumption (Figs. 2.1, 2.3 and Supplementary Figs. 2.S4–2.S8). Thus, WBA models cannot explain intergroup phonetic diversification under conditions of high intragroup variability.

The NVE model starts with the observation that vocal tract and phonetic variation is already present within every population and even exceeds between-population variation (Fig. 2.4). Moreover, the NVE model neither requires nor excludes vocal tract and/or environmental change as a condition of phonetic diversification between languages. Within the envelope of morpho-phonetic covariation, language communities are free to travel through phonetic vowel space (Fig. 2.4) in a random-like neutral manner, while deviations from neutrality are indicative of non-random phonetic changes induced by environmental and/or cultural factors.

It is often assumed that the biological "language evolution" was completed with the acquisition of the language faculty, and that all subsequent "language change" is due to socio-cultural processes (Labov, 2001). This view has recently been challenged on both methodological and empirical grounds, indicating that biological processes still influence the course of spoken language evolution (Levinson and Dediu, 2013; Blasi et al., 2019; Dediu et al., 2019). The present study adds to this evidence by showing that a key mechanism of biological evolution – within-group variation generating between-group diversity – is still at work in the phonetic diversification of languages.

2.6 Materials & Methods

Morpho-phonetic data of English speakers (MP-E). We used a subset of the USC 75-Speaker Speech MRI Database (Lim et al., 2021), which consists of video and concurrent audio recordings from Real-Time Magnetic Resonance Imaging (rtMRI) of the upper vocal tract from an open-source database of linguistically motivated speech tasks. The rtMRI data were acquired at 83.28 frames per second at an image matrix size of 84x84 pixels. Synchronous audio recordings were obtained in the scanner using fiber-optic microphones and then were subsequently denoised and are available in .wav format.

We extracted carrier words ('beat', 'bait', 'pot', 'boat', 'boot') to analyze American English vowels (i, e, a, o, u) as produced (each twice) by N=32 speakers (16 females and 16 males; aged 18–59y) from different ethnic and linguistic backgrounds. For each speaker, MRI frames were extracted from the rtMRI videos (see Supplementary Fig. 2.S1). Depending on the length of the vowel, around 10 consecutive images (corresponding to 0.12 seconds) were extracted. In total, 3210 rtMRI images were used for the morphometric analyses (32 speakers, ~100 images each across two repetitions of five vowels). The audio data were synchronized with the extracted image positions for acoustic phonetic analysis of the vowels. Thus, each extracted set of images and synchronous timestamps in the audio recordings together represent a vowel visually and acoustically.

Geometric morphometric analysis of vocal tract shape. The rtMRI images were resampled at 164x164 pixels for geometric morphometric (GM) analysis (Dryden and Mardia, 2016; Bookstein, 2018). The shape of the upper vocal tract was quantified with K=95 two-dimensional landmarks, denoting both stationary skeletal and soft tissue structures (philtrum, anterior nasal spine, upper incisor tip, hard palate, inferior end of the nasopharynx, anterior borders of intervertebral disks between vertebrae C2 to C5) and moving skeletal and soft-tissue structures (lips, chin, supramentale, tongue, soft palate, uvula, epiglottis). The landmarks consisted of anatomical reference points (K_f=21) and semilandmarks (points along curves between reference points; K_c=74). See Supplementary Fig. 2.S2 and Supplementary Table 2.S1.

Landmark positions were digitized semi-automatically on each of the 3210 rtMRI images using the R software packages *StereoMorph*, *Morpho* and *geomorph* (Olsen and Westneat, 2015; Schlager, 2017; Adams et al., 2022; R Core Team, 2024). Generalized Procrustes Analysis (GPA) (Dryden and Mardia, 2016) was carried out to scale individual landmark configurations to unit centroid size and to minimize differences due to rotation and translation of the individual configurations. GPA was performed only with those hard-tissue landmarks that remain stationary during articulation (see Supplementary Fig. 2.S2 and Supplementary Table 2.S1). For each vowel of each speaker, a mean configuration was calculated. The morphometric data matrix used for all further analyses thus consisted of $5N$=160 rows (representing the 5 vowels per N=32 individuals) and $2K$=190 columns (representing the 2D Procrustes shape coordinates of the K=95 landmarks per configuration). Principal Components Analysis (PCA) was then performed on the Procrustes shape coordinates to create a low-dimensional representation of shape space. One major advantage of GM analysis is that each articulatory shape of the vocal tract (per individual and per vowel) is represented by one multidimensional point in shape space. Conversely, each point in shape space can be re-transformed into its corresponding vocal tract shape in 2D physical space. This property facilitates the identification and visualization of major patterns of shape variation in the sample (see Figs. 1–3).

Acoustic data extraction. Vocal tract resonance frequencies (formants F1–F4) and the fundamental frequency (F0) were extracted from the audio recordings of the rtMRI videos at synchronous timestamps with the image frames. Vowel measurements were taken through visual inspection of the spectrograms using Praat (Boersma, 2001). This procedure is common practice given the unreliability of automatic formant trackers (Harrison, 2013), particularly in noisy environments (e.g., MRI scanner noise is present in higher frequency ranges).

An average length of 0.12–0.15 seconds of audio (corresponding to 10–12 video frames) was analyzed for each vowel. Vowel mid-points in utterances were used to reduce possible formant transition effects due to coarticulation. For the diphthongs 'bait' and 'boat', we extracted the formants of the first vowel (e, o) at its midpoint. Definitive vowel identification was cross-validated with the articulatory patterns in the rtMRI image frames at concurrent time stamps. Mean fundamental frequency (F0) was measured for every speaker and every vowel similarly to formant extraction. The formant measurements from the two rounds of recordings were then used to calculate a mean formant value averaged over both recordings for each vowel and each speaker.

Lastly, given the noise emitted by the MRI scanner, the acoustics of the speech recordings may contain features of the Lombard effect (Lombard, 1911; Lane and Tranel, 1971). However, since the articulatory and acoustic data are recorded simultaneously, it is reasonable to assume that both data sets are similarly biased, such that morpho-phonetic covariation remains unbiased. Nevertheless, future studies may benefit by improving the sound quality of intra-MRI audio recordings towards more natural settings.

Morpho-phonetic co-analysis. *Data preparation.* Sex and vocal tract size are known to be important factors influencing the speech signal (Simpson, 2009; Garcia et al., 2017; Munson and Babel, 2019; Serrurier and Neuschaefer-Rube, 2023). Indeed, preliminary analyses of our morpho-phonetic data showed significant effects of sex and vocal tract size on vocal tract shape and acoustic output. In this study, we focus on those patterns of interindividual articulatory and acoustic variation that are independent of sex and vocal tract size and are thus common to all individuals in the sample. To remove the effects of sex and size from the data, we proceeded as follows. The vocal tract shape data (PCs 1-10) as well as the acoustic data (formants F1-F4 and fundamental frequency F0) were regressed against centroid size of the vocal tract and sex. The respective technique (multivariate multiple regression) is analogous to a multivariate analysis of covariance (MANCOVA) and was performed with the R packages *geomorph* and *RRPP* (Collyer and Adams, 2018; Adams et al., 2022). All subsequent analyses were then performed on the residuals of the regression data.

Correlations between articulatory and phonetic distances. Correlations between articulatory and phonetic distances (Fig. 1F) were evaluated by regressing, for each individual, the N=10 Euclidean distances between the five vowels in acoustic space versus articulatory space. To compute the distances in acoustic space, we used formant wavelengths rather than frequencies. This is based on the consideration that the vocal tract acts as a resonating body, such that articulatory changes to its dimensions tend to elicit proportional changes of formant wavelengths rather than frequencies.

Covariance between articulatory and acoustic data. Both in multidimensional articulatory and acoustic space, individual vowel polygons can be characterized by three properties: a) location in space, as given by the coordinates of the polygon centroid (i.e., the average position of all vowels), b) polygon size, quantified as its centroid

size (the square root of the sum of the vowels' squared distances from the centroid); and c) polygon shape (the vowel coordinates relative to the centroid location). These geometric properties serve as proxies for the following real-life morphological and acoustic properties: in morphological terms, polygon location is a proxy of the resting state morphology of the vocal tract (the average of all vowel articulations), polygon size corresponds to the articulatory range, and polygon shape characterizes the articulatory contrasts between vowels. In acoustic terms, polygon location corresponds to the timbre of the voice, size to the acoustic range, and shape to the acoustic contrasts between vowels.

Partial Least-Squares (PLS) analysis was used to study the covariation between the articulatory and acoustic vowel polygon data, and to explore and visualize how individual vocal tract shape influences the acoustic output. PLS permits the exploration of the covariation between two blocks of multidimensional data. Here, the multidimensional data are N x $2P$ matrices of articulatory landmark data (N=number of speakers, $2P$=the 2D coordinates of P landmarks), N x $5Z$ matrices of articulatory data (N=number of speakers, $5Z$=number of vowels x number of shape PCs) or N x $5Q$ matrices of acoustic data (N=number of speakers, $5Q$=number of vowels x number of formants). We used PLS to evaluate the covariation between palate shape (P_p=31) and tongue shape (P_t=50) (Fig. 2), and between the vowel polygons in articulatory space ($5Z$=50) and in acoustic space ($5Q$=10) (Fig. 3). Similar to PCA, PLS is a dimension reduction technique. It uses singular value decomposition of the covariance matrix of two data blocks (i.e., data matrices) to extract pairs of orthogonal components (termed here PLS component pairs) that capture the largest, second-largest, etc., proportions of the total covariance in the sample. Due to the high dimensionality of the data, significance levels for PLS covariance (p_{pls}) were estimated empirically with the permutation procedures (1000 iterations) implemented in the R package *geomorph* (Adams et al., 2022). We also evaluated the covariance ratio (CR) (Adams, 2016), which is a dimension-independent measure of the degree of association between two sets of multivariate data. The CR quantifies the covariance between two multidimensional data blocks as a fraction of the covariance within each block.

Data visualization and rendering. To visualize the correlated articulatory-acoustic (morpho-phonetic) variation, we proceeded as follows. For Figure 2, we evaluated the average tongue and palate shape of the three most extreme individuals at opposite ends of the data scatter (Fig. 2A). To generate the vocal tract profiles (Fig. 2C) of the two extremes, we calculated the distances between 21 tongue landmarks (tongue tip to tongue back) and the palate outline. Distances were measured perpendicular to the central axis of the air tube using ImageJ (Schneider et al., 2012). A similar approach was used for Figure 3, where the average vocal tract and formant configurations of each three individuals at the opposite ends of the data scatter were calculated (Fig. 3A), and visualized in Figs. 3A–C.

Cross-linguistic phonetic data (VC-C). In order to compare the patterns of group-level morpho-phonetic variation in the MP-E sample with patterns of language-level variation, we used the VoxCommunis corpus (Ahn and Chodroff, 2022) (VC-C), a large-scale cross-linguistic data set, which builds on a massively-multilingual collection of transcribed speech from the publicly available Mozilla Common Voice Corpus (Ardila et al., 2020). The VC-C dataset includes acoustic-phonetic measurements (F1-F4) of 36 different languages. From these, we selected the languages that included the five cardinal vowels (i, e, a, o, u) that we used in our morpho-phonetic

analyses of the MP-E data. We excluded data with obvious errors or too few speakers. We analyzed data points only from speakers who produced the full set of vowels (i, e, a, o, u), in order to get the full vowel polygon of each speaker. This led to a sample of 19,335 data points (vowel utterances), comprised of 3867 individuals from 10 different languages (Basque, Bulgarian, Hindi, Italian, Portuguese, Romanian, Russian, Tamil, Turkish, Vietnamese). To directly compare the VC-C data with our MP-E data, we adjusted the data for differences in formant extraction methods by superimposing vowel centroids. Furthermore, we ran similar data adjustment procedures as for the MP-E in order to obtain sex-adjusted between-language data (pooled for sex, N=19,335). Variation in the sex-pooled phonetic data was then decomposed into variation within languages ($V_{\text{pho.w}}$), i.e., shared by all languages and variation between languages ($V_{\text{pho.b}}$). $V_{\text{pho.w}}$ was evaluated by pooling the data for language (N=19,335), while $V_{\text{pho.b}}$ was evaluated from the mean data per language.

Comparing patterns of variation. *Graphical comparison.* We compared the principal patterns of phonetic variation in the MP-E sample and in the VC-C sample (both within-languages and between-languages) by projecting all data into formant space (Fig. 4A; variation per vowel) as well as into a common PC space (Fig. 4B; variation of the vowel polygon). To construct the graph of Fig. 4B, the vowel polygon data were expressed as data vectors of length $5Q$ (5 vowels, Q=2 formants). PCA was performed on the 10x10 data matrix of the mean vowel polygons of the 10 languages, such that PC1 of Fig. 4B expresses the principal direction of between-language variation of the vowel polygon. The principal directions of within-language variation in the VC-C and MP-E samples were projected into the same PC space. Finally, the principal direction of morpho-phonetic covariation in the MP-E sample (acoustic PLS1 in Fig. 3A) was also projected into the PC space.

Statistical comparison of variance-covariance structure. The near-collinearity of the principal variation vectors in Fig. 4B indicates proportionality of the variance-covariance matrices underlying these distributions. Statistical tests on proportionality were performed with eigenanalysis (Bookstein and Mitteroecker, 2014), as implemented in the R package *vcvcomp* (Le Maître and Mitteroecker, 2019) (see *Vocal tract variation influences within- and between-language phonetic diversity* in main text).

Test for neutral vowel evolution. In analogy to the standard model of neutral phenotypic evolution, we expect that, under neutral conditions of language change, the pattern of vowel polygon variation between languages is proportional to the variation within languages, $V_{\text{phon.b}} \sim V_{\text{phon.w}}$ (Lande, 1980). This proposition can be tested as follows (Ackermann and Cheverud, 2004): the VCV (variance-covariance) matrices characterizing between-language and within-language variation are each orthogonalized, yielding statistically independent variance components along the eigenvectors. Then the between-group variance components are regressed against the within-group components. Proportionality between the two VCV matrices is indicated by linearity (i.e., slope $= 1$ in a double-logarithmic plot; Supplementary Fig. 2.S9).

2.7 Supplementary Information

This supplementary Information includes supplementary Figures S1–S9 and supplementary Tables S1–S2.

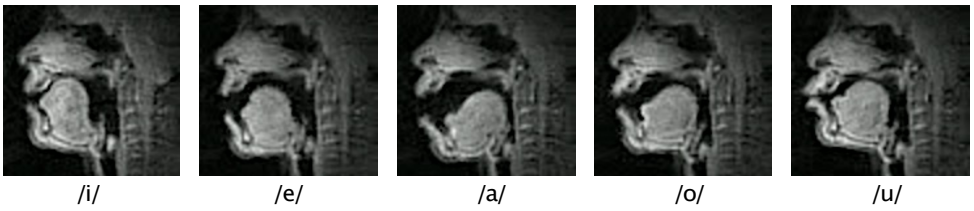
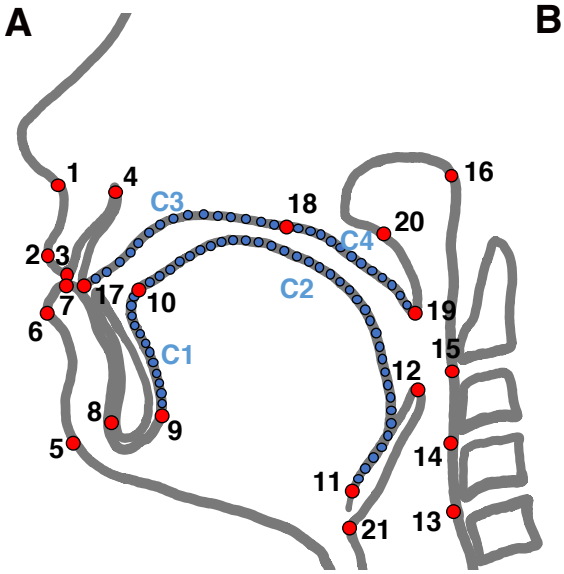


FIGURE 2.S1: **Example of rtMRI image frames during vowel production.**



B

NR	Abbr. / Description
1	P / Philtrum
2	AUL / Anteriormost point of upper lip
3	IUL / Inferiormost point on upper lip
4	ANS / Anterior nasal spine
5	MP / Mental protuberance
6	ALL / Anteriormost point on lower lip
7	ILL / Inferiormost point on lower lip
8	SM / Supramentale
9	TB / Anterior base of tongue
10	TT / Tongue tip
11	TR / Tongue root
12	ET / Epiglottis tip
13	ID3 / Intervertebral disc C5/C4
14	ID2 / Intervertebral disc C4/C3
15	ID1 / Intervertebral disc C3/C2
16	IEN / Inferior end of nasopharynx
17	FT / Tips of upper incisors
18	BPA / Border hard-soft palate
19	VT / Velum tip
20	VUB / Velum upper bend
21	ER / Epiglottis root
C1	TB-TT / Tongue base to tip (12)
C2	TT-TR / Tongue tip to root (35)
C3	FT-BPA / Outline of the hard palate (15)
C4	BPA-VT / Outline of the soft palate (12)

FIGURE 2.S2: **Schematic representation of landmarks placed on 2D midsagittal images of the vocal tract. A**: Anatomical landmarks (reference points) are in red and semilandmarks in blue. **B**: Landmark definitions. See Extended Data Table 2.S1 for detailed descriptions of the location of the landmarks.

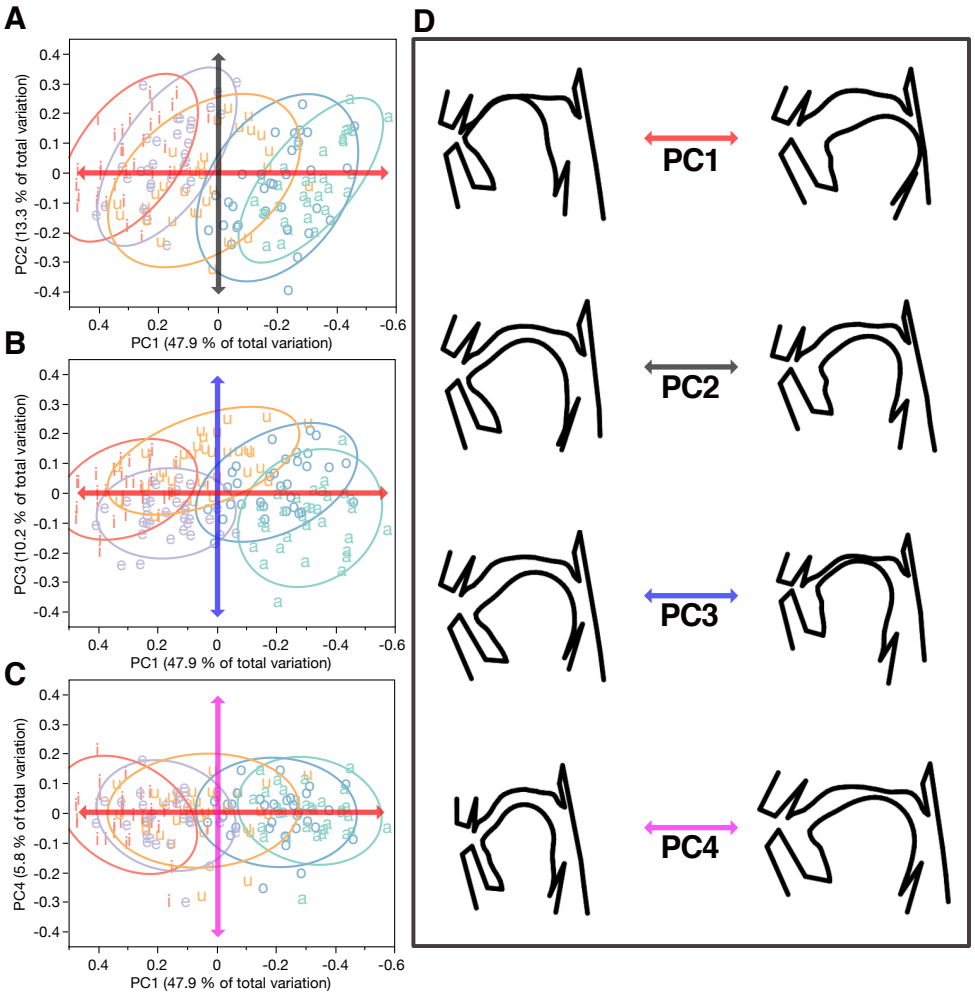


FIGURE 2.S3: **Articulatory shape variation associated with the PCs of shape space**. **A**: Shape variation along PCs 1 and 2. **B**: Shape variation along PCs 1 and 3. **C**: Shape variation along PCs 1 and 4. **D**: Visualization of vocal tract shape variation along PCs 1-4 at the two ends of the respective data scatter.

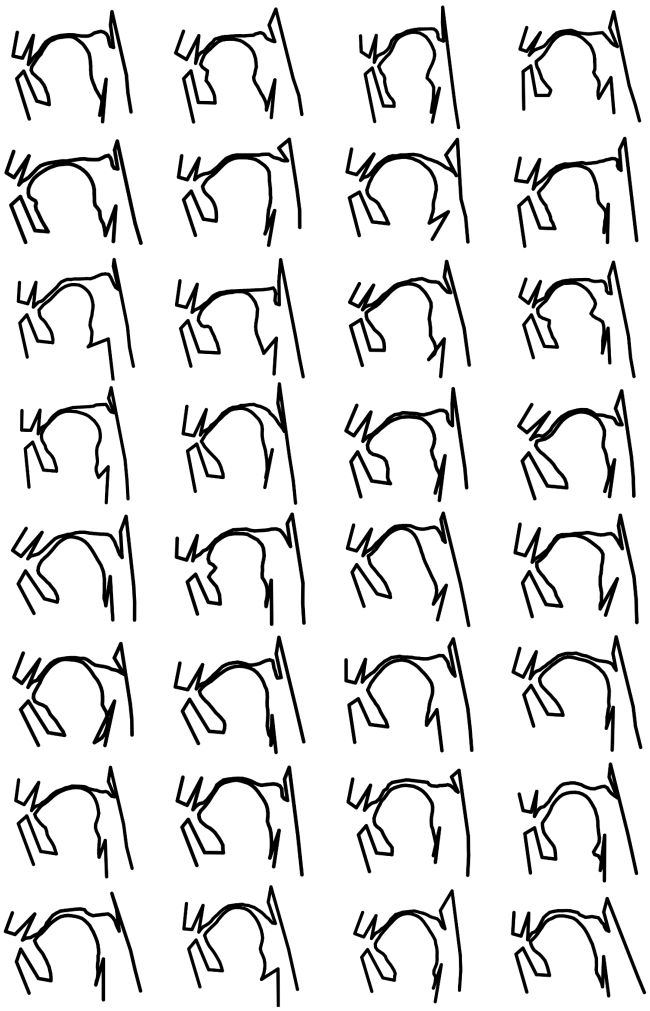


FIGURE 2.S4: **Physical shape variation in the full sample of speakers producing the vowel [i].**

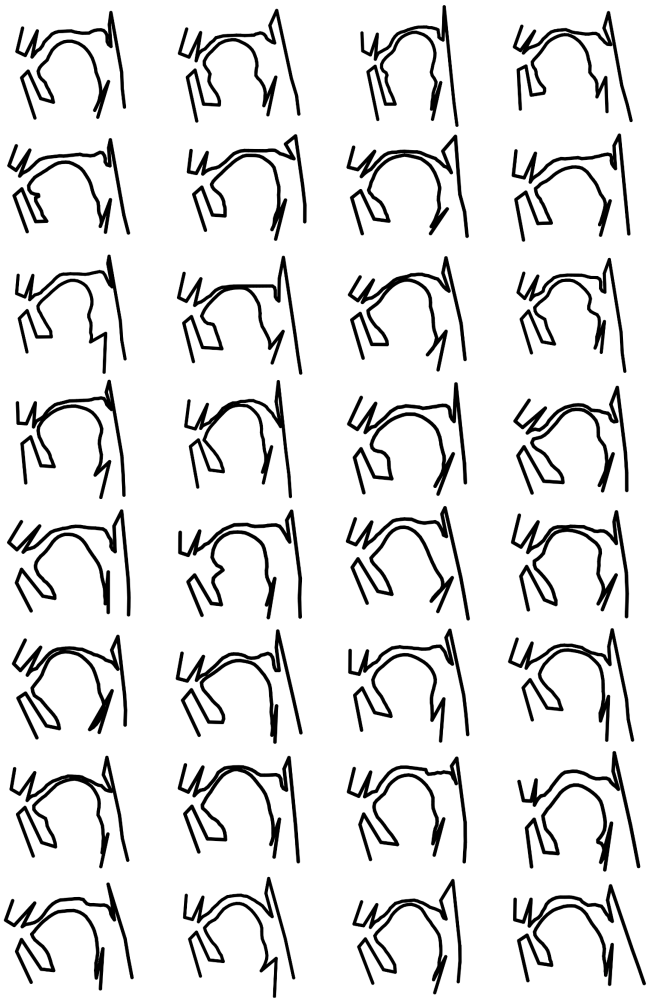


FIGURE 2.S5: **Physical shape variation in the full sample of speakers producing the vowel [e].**

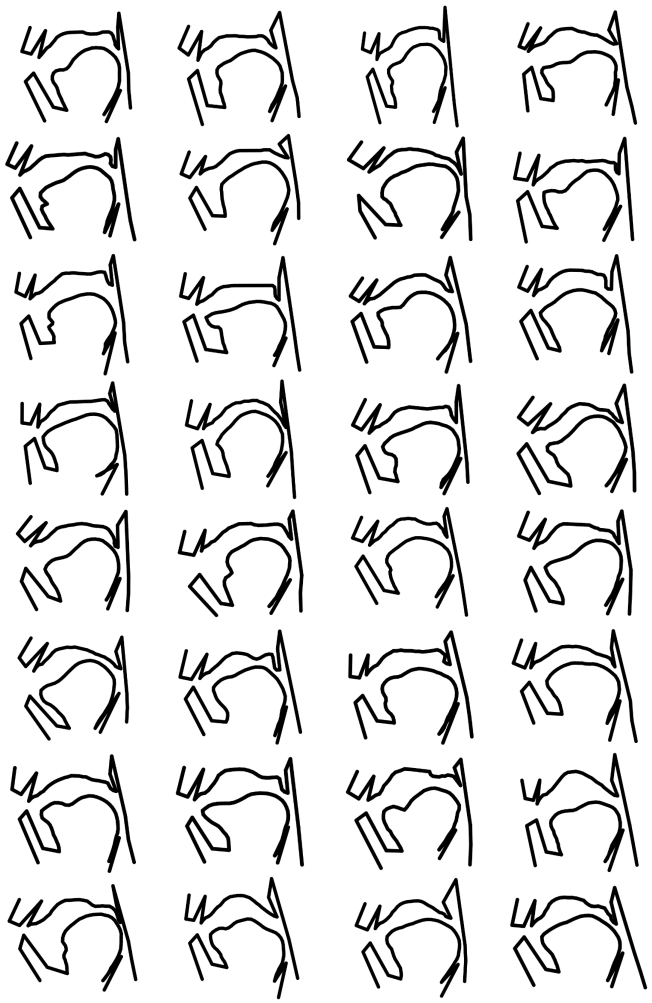


FIGURE 2.S6: **Physical shape variation in the full sample of speakers producing the vowel [a].**

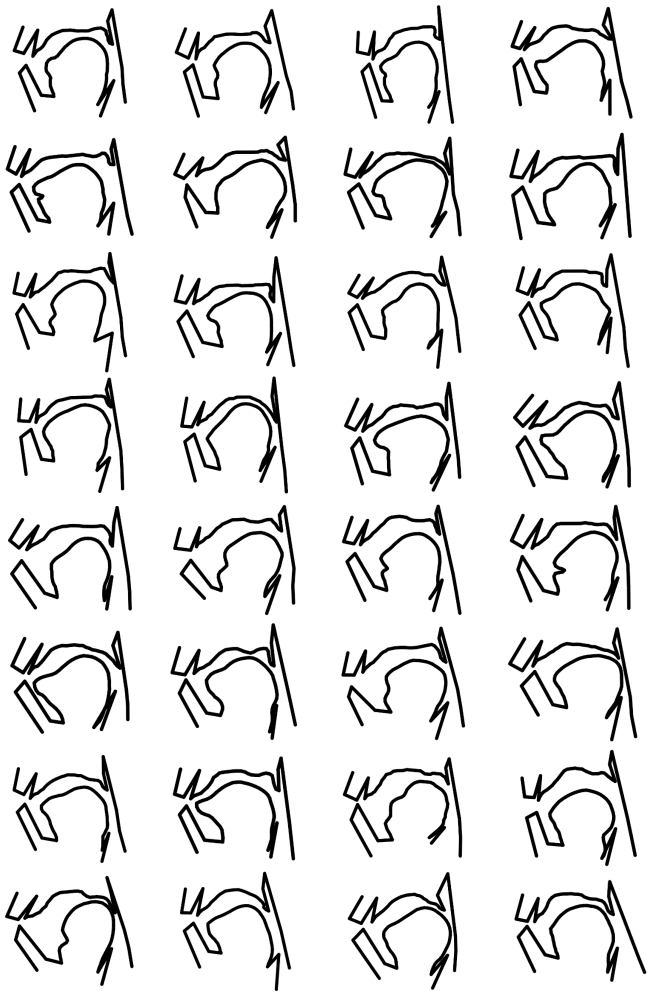


FIGURE 2.S7: **Physical shape variation in the full sample of speakers producing the vowel [o].**

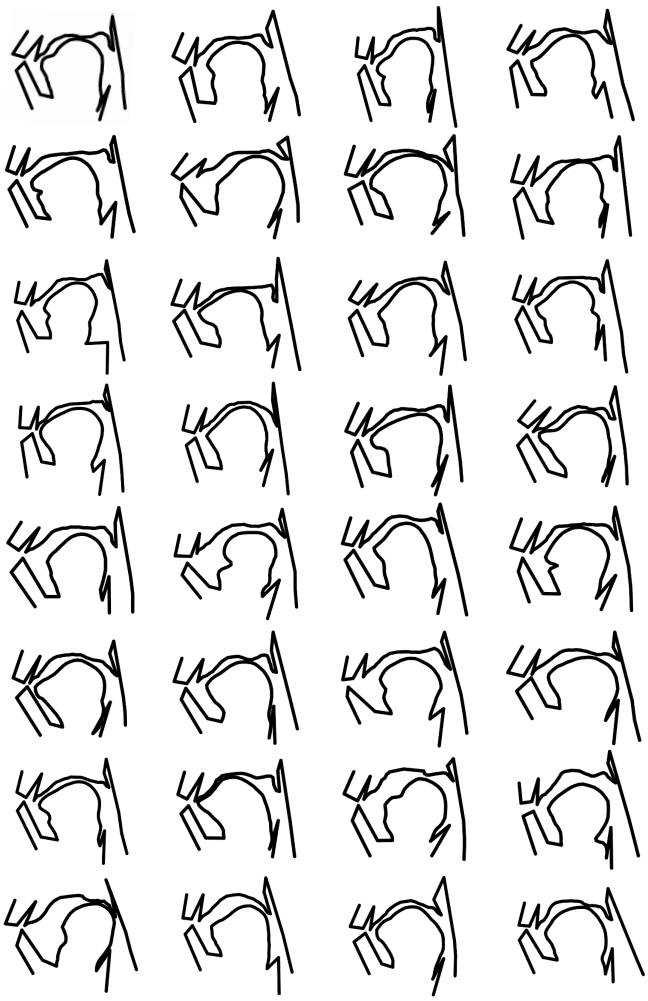


FIGURE 2.S8: **Physical shape variation in the full sample of speakers producing the vowel [u].**

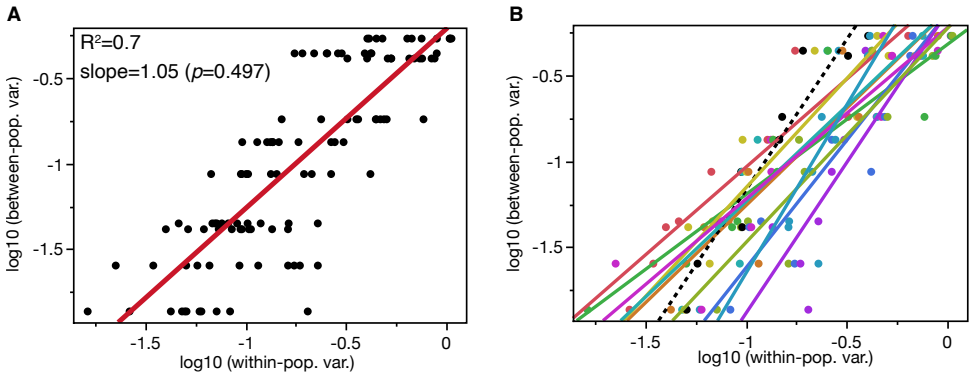


FIGURE 2.S9: **Regression of between-population (between-language) variances on within-population variances. A**: Regression of between-language variance on within-population variance by means of PCs 1-10 of vowel polygon form in F1-F2 formant space. The slope does not deviate significantly from linearity (p=0.497) The within-population variances were computed on the MP-E data and the VC-C data. **B**: Separate regression for each VC-C language (each color corresponds to a different language) and the MP-E (black dashed line).

TABLE 2.S1: **Orofacial landmarks (LM) used in this study.**

ID	Code	Description	K	Type	Tissue
1	P	Philtrum	1	F	H
2	AUL	Anteriormost point on upper lip	1	F	S
3	IUL	Inferiormost point on upper lip	1	F	S
4	ANS	Anterior nasal spine	1	F	H
5	MP	Mental protuberance	1	F	H
6	ALL	Anteriormost point on lower lip	1	F	S
7	ILL	Inferiormost point on lower lip	1	F	S
8	SM	Supramentale	1	F	H
9	TB	Anterior base of tongue	1	F	S
10	TT	Tongue tip	1	F	S
11	TR	Tongue root	1	F	S
12	ET	Epiglottis tip	1	F	S
13	ID3	Intervertebral disc C5/C4	1	F	H
14	ID2	Intervertebral disc C4/C3	1	F	H
15	ID1	Intervertebral disc C3/C2	1	F	H
16	IEN	Inferior end of nasopharynx	1	F	H
17	FT	Tips of upper incisors	1	F	H
18	BPA	Border hard-soft palate	1	F	S/H
19	VT	Velum tip	1	F	S
20	VUB	Velum upper bend	1	F	S
21	ER	Epiglottis root	1	F	S
Curve1	TB - TT	Tongue base to tip	12	C	S
Curve2	TT - TR	Tongue tip to root	35	C	S
Curve3	FT - BPA	Outline of the hard palate	15	C	H
Curve2	BPA - VT	Outline of the soft palate	12	C	S

Landmark identifiers (ID) as in Extended Data Fig. 2.S2. K: Nr. of semilandmarks per curve; Type: fixed landmark (F), curve semilandmark (C); Tissue: soft (S), hard (H).

TABLE 2.S2: **Variance ratios of within-vowel variance vs. between-vowel variance.**

Vowel	$\mathbf{VR^{Articulatory}}$	$\mathbf{VR^{Formant}}$
a	0.546	0.064
e	0.501	0.040
i	0.495	0.053
o	0.585	0.087
u	0.611	0.079
Mean	**0.548**	**0.065**

Coarticulation preserves relative articulatory and acoustic contrasts between successive phonemes

3.1 Context

The following manuscript with the title *Coarticulation preserves relative articulatory and acoustic contrasts between successive phonemes* is currently in preparation for submission to a scientific journal. This study was conducted by the author in collaboration with Steven Moran, Daniel Friedrichs, Marcia Ponce de León and Christoph Zollikofer.

3.2 Abstract

Using geometric morphometrics and synchronized acoustic analysis of real-time MRI data, we examined how coarticulation influences articulation and acoustics in vowel-consonant-vowel (VCV) sequences. In addition to introducing a methodology for the integrated analysis of articulatory and acoustic data, our findings confirm that consonants and vowels shift in articulatory and acoustic space, but preserve their relative locations, highlighting the fundamental role of contrast maintenance in continuous speech. These results underscore that speech production emphasizes relational rather than absolute phone positions. The methodology and insights presented here offer a fresh perspective on the complex mechanisms underpinning (co)articulation.

3.3 Introduction

Coarticulation, the overlapping of articulatory movements of adjacent speech sounds, is a ubiquitous feature of fluent speech. It causes phonemes to be realized with context-dependent variability: articulatory gestures for one sound are influenced by neighboring sounds, often making those sounds more similar to each other. While coarticulation contributes to the smoothness and efficiency of speech, it also poses a risk to phonemic contrast. However, loss of contrast in everyday speech is rare, implying that speakers employ mechanisms to prevent coarticulation from erasing important differences between phonological categories (Tilsen, 2013). This raises the question of how speakers balance the natural tendency for articulatory overlap with the necessity of maintaining clear, distinguishable sound units for listener comprehension.

Coarticulation has been extensively studied from the perspectives of neuromotor control, articulatory biomechanics, and acoustic output (Munhall et al., 2000; Viswanathan et al., 2010; Lindblom and MacNeilage, 2011; Gick et al., 2013). Various models seek to explain why and how phonemes influence each other during speech production and perception. For example, Lindblom (Lindblom, 1963) and Öhman

(Öhman, 1966; Öhman, 1967) proposed a principal difference between the stationary, isolated expression of "pure" phonemes, and their dynamic expression as part of a phoneme sequence. Lindblom's centralization model attributes reduced vowel space in stationary vowels to "target undershoot", implying that dynamically spoken vowels do not always attain their stationary target locations in formant space, and/or their target shapes in terms of articulatory morphology (Lindblom, 1963). In comparison, Öhman's coarticulation model proposes that phonemes shift articulatorily and phonetically toward their neighbors (Öhman, 1967). Thus, although coarticulation can enhance fluidity, it also needs to be constrained so that important contrasts are not neutralized (Tilsen, 2013).

In practical terms, coarticulation is often measured by changes in the dynamic properties of a phoneme as a function of its phonetic context (Öhman, 1966; Öhman, 1967). Here, we quantify coarticulation by tracking shape changes in vocal tract morphology and analyzing the acoustic signal at synchronous time stamps. Articulatory gestures are captured via real-time Magnetic Resonance Imaging (rtMRI) data from speakers producing vowel-consonant-vowel (VCV) triplets (Sorensen et al., 2017). The shape of the upper vocal tract is quantified using geometric morphometrics (GM; Bookstein, 2018) with a set of $N = 172$ anatomical landmarks (Fig. 3.1a). We analyze speaker acoustics by measuring Mel-Frequency Cepstral Coefficients (MFCCs) and formants (F1 & F2; see Fig. 3.1b,c). To the best of our knowledge, the full set of GM tools combined with acoustic measures has not yet been applied to investigate coarticulation in production and perception (cf., Zollikofer et al., 2011; Gully, 2021; Moran et al., 2024). GM allows for a detailed representation of coarticulatory variation in both multivariate shape space and the physical vocal tract space. Importantly, this model-free, "morpho-phonetic" approach does not presuppose idealized phonetic targets, but rather uses the average acoustic output as its frame of reference.

Given these methods, our broader aim is to explore how speakers manage coarticulation without obliterating phonemic distinctions. By integrating GM-based articulatory analyses with acoustic parameter tracking, we examine whether – and to what extent – phonemes influence each other in continuous speech, and how speakers maintain sufficient contrast so that each phoneme remains perceptually salient.

3.4 Materials and Methods

3.4.1 Data

The data used in this study come from the publicly available *USC Speech and Vocal Tract Morphology MRI Database* (Sorensen et al., 2017). Four native American English speakers (two women and two men) produced the vowel-consonant-vowel (VCV) sequences [iCi, aCa, uCu]. Each VCV sequence consists of a middle consonant (C), represented by the consonants [p, t, k, s, l], and the surrounding initial (v_1) and final (v_2) vowel, represented by the three cardinal vowels [i, a, u]. These sequences allow us to investigate effects of both anticipatory and perseverative coarticulation. The midsagittal rtMRI data were recorded at a rate of 23.18 frames per second (fps; 43.13 ms interval), and at a matrix size of 60x60 pixels. Simultaneously, time-aligned and denoised audio data were acquired with the rtMRI frames within the MRI scanner. Details on the collection of the audio recordings are available in Sorensen et al. (2017). In total, $N = 891$ data points, i.e., rtMRI frames and corresponding acoustic time stamps were collected.

3.4.2 Geometric morphometric analysis

The shape of the upper vocal tract was quantified throughout the VCV-sequences by placing $N = 172$ landmarks on the rtMRI frames. The landmarks consist of $K_f = 20$ two-dimensional fixed as well as $K_c = 152$ semi-landmarks along curves (denoting both stationary skeletal structures and moving soft-tissue structures) on the rtMRI images (Fig. 3.1a). Landmark positions were digitized semi-automatically using the R programming language and the *StereoMorph* library (Olsen and Westneat, 2015; R Core Team, 2024). The resulting morphological data set consists of $N = 891$ landmark configurations, each of which is represented by the xy-coordinates of the $N = 172$ 2D-landmarks.

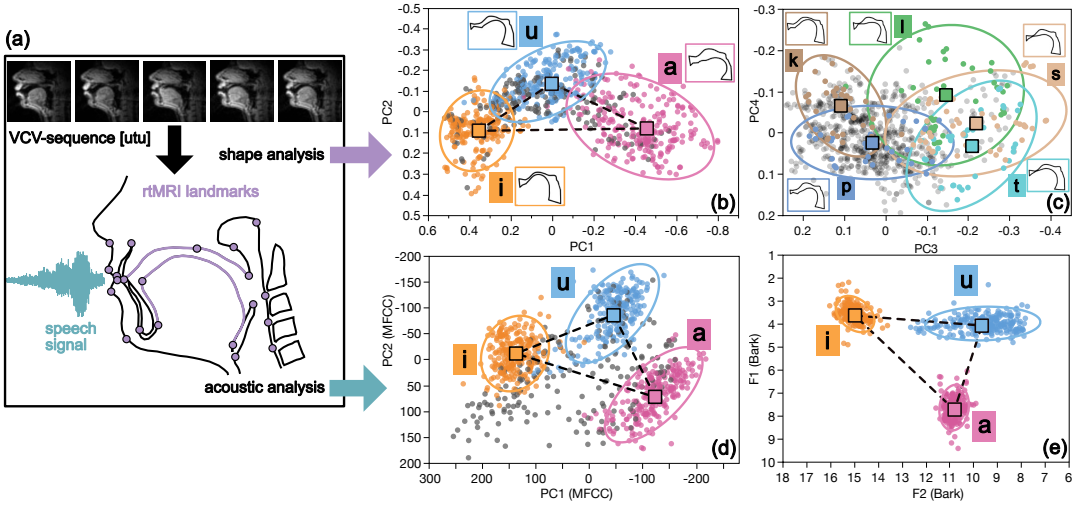


FIGURE 3.1: **Co-analysis of articulatory and acoustic variation in VCV sequences**. **(a)** Articulatory gestures are quantified by locating N=172 anatomical landmarks on each rtMRI frame of an utterance (example: five frames of the sequence utu). The corresponding acoustic output is quantified by extracting MFCCs and formants. **(b)** – **(c)**: Geometric-morphometric analysis of articulatory variation. Principal components 1 and 2 capture variation in vowel gestures (**b**). Density ellipses (90%) correspond to each vowel subspace. The vowel means (squares) form the articulatory vowel triangle. The inset graphs represent shape renderings of each mean-vowel gesture. The gray points represent coarticulated consonants. PCs 3 and 4 capture variation in consonant gestures (**c**). Density ellipses (90%) correspond to each consonant subspace. The gray points represent coarticulated vowels. The inset graphs represent shape renderings of each mean-consonant gesture. **(d)** – **(e)**: Acoustic variation in MFCC PCA space **(d)** and in Bark scaled formant space **(e)**. Colored points each correspond to a vowel and the coarticulated consonants (gray points). Squares indicate mean vowel acoustics and together form the acoustic vowel triangle. Density ellipses (90%) correspond to each vowel subspace.

Geometric morphometric (GM) analysis was carried out on the landmark configurations according to standard procedures using Generalized Procrustes Analysis (GPA), principal components analysis (PCA) of the multivariate GPA data, and visualization of patterns of shape variation in both PCA and physical spaces (Dryden and Mardia, 2016; Bookstein, 2018; see Figs. 3.1b–c). All geometric morphometric procedures were computed using the R library *geomorph* (Adams et al., 2022). GM permits the simultaneous representation of morphological data in real physical space (the actual articulatory gestures) and in multivariate shape space (defined by shape PCs). These spaces are best suited for the visual representation and the

statistical analysis of the data, respectively. Accordingly, the specific landmark configuration of the articulatory gesture at any given point in time corresponds to a specific location in multidimensional shape space. Reciprocally, any point in shape space can be retransformed into the specific configuration of the articulatory system (Figs. 3.1c–e). Note that for the visualization of the articulatory gestures in real physical space, we chose only a selection of landmarks that represent the vocal tract airway contours during articulation (Figs. 3.1–3.3).

3.4.3 Acoustic analysis

The acoustic analyses focused on two sets of acoustic parameters: Mel-Frequency Cepstral Coefficients (MFCCs; Davis and Mermelstein, 1980) and vowel formants. While formant tracking is appropriate to capture vowel resonances, consonantal acoustics often involve complex or noise-like spectra that do not lend themselves to clear formant measurement. Thus, we used MFCCs to capture broader spectral information across both consonants and vowels in a unified space. We measured the first two formant frequencies (F1 & F2) in Praat (version 6.4.23; Boersma, 2001), adjusting analysis parameters for adult male vs. female voices to account for typical sex-based differences in vocal tract length. Analysis settings followed typical formant-tracking approaches. We then converted the resulting F1–F2 values to Bark scale (Traunmüller, 1990) for plotting and interpreting the acoustic–perceptual vowel space (Fig. 3.1c). From the same acoustic frames, we computed 12 MFCCs (excluding the zeroth coefficient) in Praat, using 15ms windows with a 43.13ms frame shift, followed by standard liftering and mean normalization. We performed Principal Components Analysis (PCA) on these 12 MFCC vectors, retaining the first two principal components (MFCC-PCs) for low-dimensional visualization of acoustic–perceptual variation of vowels and consonants (Fig. 3.1b). Both feature sets (formant frequencies and MFCCs) were measured at the same timestamps as the rtMRI frames (i.e., each 43.13ms), ensuring that each rtMRI frame has temporally aligned acoustic observations for subsequent morpho-phonetic co-analysis (see Section 3.4.2). Ultimately, by combining acoustic analysis (Section 3.4.3) with GM analysis (Section 3.4.2), i.e., using a "morpho-phonetic" approach, we establish a quantitative multivariate correspondence between articulatory gestures and their acoustic outcomes (Figs. 3.1b, d–e), which permits us to test various hypotheses regarding coarticulatory effects in both domains.

3.4.4 Statistical analysis

Because sex and vocal tract size systematically affect both articulation and acoustics (Simpson, 2009; Garcia et al., 2017; Munson and Babel, 2019; Serrurier and Neuschaefer-Rube, 2023), we analyzed all coarticulatory effects by controlling for these factors. Technically, this was achieved by regressing the multivariate articulatory and acoustic data on sex and vocal tract size, and computing all further analyses on the residuals of the regression data (see Figs. 3.1b–e).

We analyzed coarticulatory interactions between neighboring phones in both articulatory (shape) and acoustic (formant) space, focusing on two primary directions of influence: the effect of consonants on vowels, and the effect of vowels on consonants (Figs. 3.2 & 3.3). These interactions were examined both visually and statistically. To visualize coarticulatory effects, we plotted articulatory and acoustic variation in the respective multivariate spaces (Figs. 3.2 & 3.3). To statistically assess these

effects, we applied multiple multivariate regressions using the R packages *geomorph* and *RRPP* (Collyer and Adams, 2018; Adams et al., 2022).

We tested coarticulatory effects on multivariate vowel data as a function of three factors: *phoneme*, *consonant context*, and *vowel position* (Table 3.1). The *phoneme* factor captures the overall differences between the three vowels [i, u, a] (Fig. 3.1b, d–e). The *vowel position* factor refers to the effect of the initial (v_1) vs. final (v_2) vowel in the VCV sequence (Fig. 3.2a–c), while the *consonant context* factor reflects the influence of the intervocalic consonant on the adjacent vowels (see Fig. 3.2d–g).

Similarly, we assessed coarticulatory effects on consonant production using two predictors: *phoneme*, representing differences among the consonants [p, t, k, s, l] (Fig. 3.1c), and *vowel context*, which captures the influence of the surrounding vowels on the consonant within each VCV sequence (Table 3.2; Fig. 3.3).

Moreover, morphological and acoustic differences between adjacent segments were quantified using Euclidean distance in multivariate space. In shape space, this reflects the amount of articulatory change; in MFCC space, this reflects the amount of acoustic change. By comparing these distances, we assessed how gestural coarticulation translates into acoustic effects. Specifically, we measured articulatory and acoustic distances between consecutive segments in all VCV sequences (Fig. 3.4a). We also examined how vowel-consonant distances relate to coarticulatory magnitude, defined as the distance between a phoneme in context and its mean position (Fig. 3.4b). This allowed us to test whether coarticulatory effects depend on segmental proximity.

3.5 Results

3.5.1 Articulatory and acoustic variation

Figures 3.1b–c show the articulatory variation of vowels and consonants in shape space, based on the first few shape PCs. Each point in the shape space plot corresponds to a specific landmark configuration. PC1 and PC2 capture the principal shape differences between vowels, while higher-order PCs (PC3 & PC4) more specifically distinguish between consonants (see Fig. 3.1c). Figures 3.1d–e illustrate the corresponding acoustic variation of vowels and consonants in all VCV-sequences in multivariate MFCC space and in formant space. The mean vowel positions form the classic acoustic vowel-triangle. Notably, multivariate MFCC and formant spaces are remarkably similar to the articulatory space. While vowels occupy relatively distinct regions in articulatory and acoustic space, consonants show considerable overlap in their multivariate distributions in both articulatory and acoustic spaces.

	Articulation			**Formants**			**MFCCs**		
Factor	R^2	Z	p-value	R^2	Z	p-value	R^2	Z	p-value
Phoneme [i, a, u]	0.70174	8.0612	0.001	0.87705	16.52	0.001	0.67129	9.07	0.001
Consonant context [p, t, k, s, l]	0.06031	13.69	0.001	0.00244	3.54	0.001	0.01544	7.02	0.001
Vowel position [v_1, v_2]	0.00167	2.73	0.001	0.00265	4.48	0.001	0.00532	5.16	0.001
Residuals	0.23629	-	-	0.11786	-	-	0.30795	-	-

TABLE 3.1: Variance explained (R^2), effect sizes (Z), and significance (p-values) for the effects of different factors on single vowel locations in articulatory space, formant space, and MFCC space using multiple multivariate regression. The p-values represent empirically computed p-values over 1000 random models.

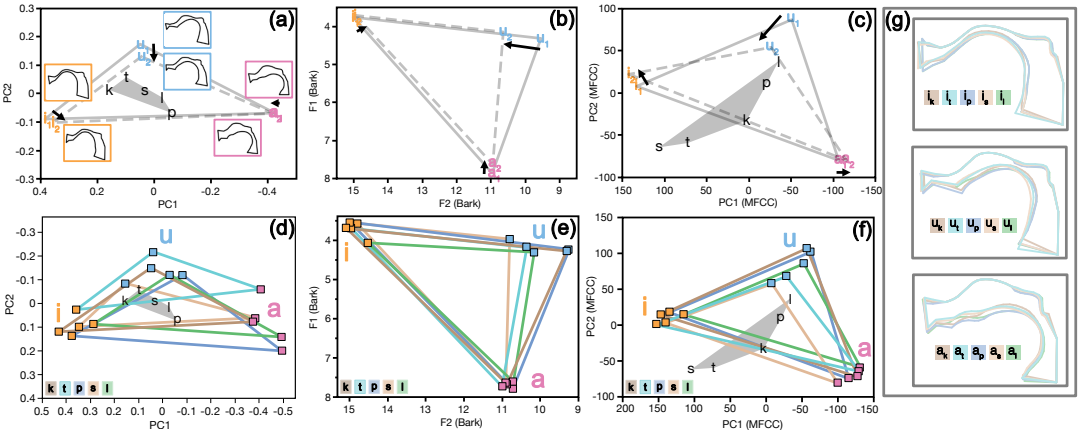


FIGURE 3.2: **Coarticulatory effects on vowels. (a) – (c)** Average articulatory and acoustic shifts of initial [i_1, u_1, a_1] vs. final [i_2, u_2, a_2] vowels. Overall vowel shifts are indicated with black arrows, and the change in the mean vowel triangle in articulatory shape space **(a)**, formant **(b)** and MFCC-PC **(c)** vowel space (initial solid gray, final dashed gray). Mean locations of consonants are shown in articulatory shape space and MFCC-PC acoustic space in black. **(d) – (f)** Effect of coarticulation on mean articulatory **(d)**, formant **(e)** and MFCC-PC **(f)** vowel triangles in different consonant contexts. Black consonants refer to the mean consonant locations. **(g)** Renderings of average vowel gestures in different consonant contexts.

3.5.2 Coarticulatory effects on vowels and consonants

Coarticulation in VCV sequences leads to relative shifts of the production of the initial and final vowels, and the overall vowel triangle in both articulatory and acoustic spaces. Figures 3.2a–c show that the shifts between initial and final vowels are moderate on average, although the differences are statistically significant (Table 3.1). Figures 3.2d–f show that the vowel triangle is pulled toward each consonant's region for the distinct consonant contexts, and vowels coarticulated with alveolars [t, s, l] exhibit a slightly reduced acoustic range. This pulling effect leads to slightly altered articulatory vowel gestures for each consonant context (Fig. 3.2g). Overall, plosives [p, k] tend to produce smaller shifts in vowels than alveolars [t, s, l]. These results indicate that consonant context influences both within- and between-vowel variation.

Adjacent vowels significantly affect consonant articulation and acoustics (Fig. 3.3 & Table 3.2). Figures 3.3a–b show that the location of each consonant varies widely across vowel contexts in articulatory shape space, indicating that consonant gestures are "pulled" toward the vowel's articulatory configuration. A similar effect is observed in acoustic space (Fig. 3.3c–d). This observation suggests that consonants are only weakly constrained by their intended articulatory target, i.e., the consonant's typical position in the absence of strong coarticulatory influence. An examination of the actual articulatory gestures (Fig. 3.3e) confirms that tongue posture is strongly influenced by the adjacent vowel, while the essential articulatory constriction for each consonant remains stable (e.g., alveolar contact in [t, s]).

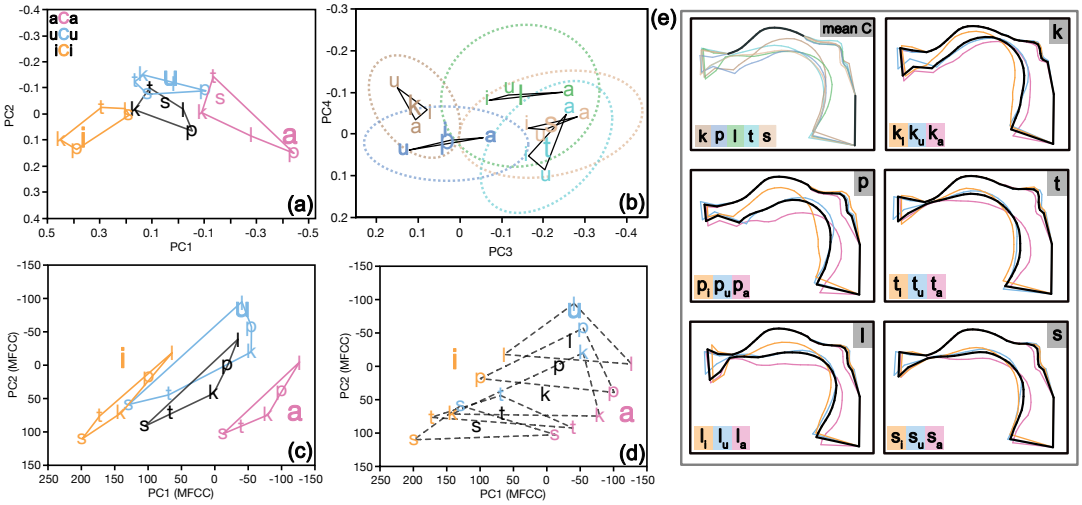


FIGURE 3.3: **Coarticulatory effects on consonants. (a) – (b)** Coarticulatory effects for consonants in articulatory shape space. The consonant polygon (k–p–l–s–t) shows vowel-associated articulatory shifts (colored according to vowel-context). The black polygon indicates the mean locations of the consonants. Mean vowels are colored accordingly. The density ellipses in **(b)** show the overall variation of consonants in PC3–PC4 articulatory shape space around the mean consonant, whereas the mean vowel-associated consonant for each vowel context is indicated (consonant triangles). **(c) – (d)** Coarticulatory effects for consonants in acoustic space. The consonant polygons in **(c)** show the vowel-associated consonant subspaces. The consonant triangles in **(d)** show the respective vowel-associated acoustic pull of consonants (colored accordingly). **(e)** Coarticulatory effects on consonants in physical shape space. The shape renderings correspond to subgraphs **(a) – (b)** and visualize the mean consonants associated with each vowel context. The top left panel shows the average consonant gestures.

	Articulation			**MFCCs**		
Factor	R^2	Z	p-value	R^2	Z	p-value
Phoneme [p, t, k, s, l]	0.27427	8.01	0.001	0.28748	9.08	0.001
Vowel context [i, a, u]	0.42057	7.67	0.001	0.31743	6.19	0.001
Residuals	0.30516	-	-	0.39510	-	-

TABLE 3.2: Variance explained (R^2), effect sizes (Z), and significance (p-values) for the effects of different factors on single consonant locations in articulatory space and MFCC space using multiple multivariate regression. The p-values represent empirically computed p-values over 1000 random models.

3.5.3 The effect of articulatory distance on coarticulation

Acoustic trajectory length – measured as the summed distance between consecutive phonemes in a VCV sequence – correlates significantly with articulatory trajectory length (Fig. 3.4a), suggesting that gestural coarticulation translates into the acoustic domain. We also examined whether the articulatory distance between vowels and consonants of a given VCV sequence predicts coarticulatory magnitude, defined as the vowel's shift from its mean position in a given consonant context. Results show a positive correlation: larger vowel-consonant distances lead to greater vowel shifts in both articulatory and acoustic space (Fig. 3.4b). In contrast, consonant coarticulatory magnitude does not depend on vowel-consonant distance ($p > 0.5$). This likely reflects the tighter biomechanical and aerodynamic constraints on consonants – especially plosives and fricatives – which must maintain precise constrictions to preserve their identity.

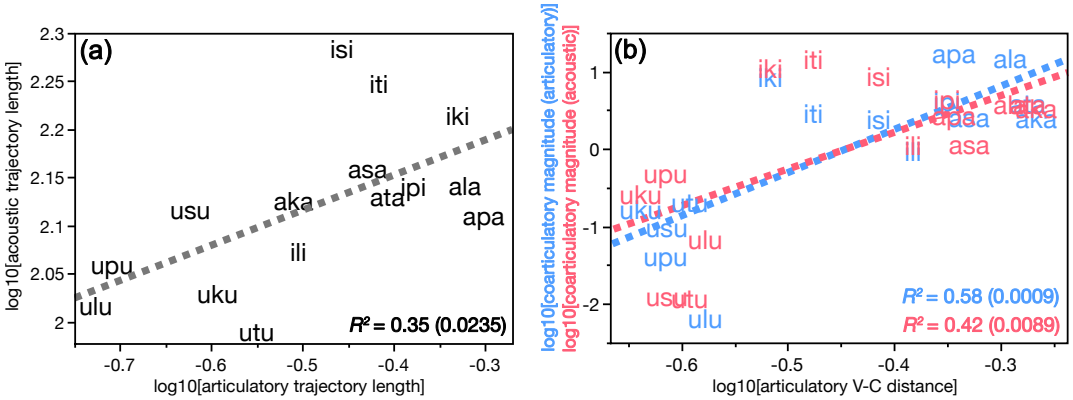


FIGURE 3.4: **Correlation of articulatory-acoustic distances. (a)** Log-log plot showing the relationship between articulatory and acoustic trajectory lengths (the cumulative distances traveled from V to C to V, in articulatory and acoustic spaces, respectively). The positive correlation suggests that larger articulatory changes correspond to greater acoustic differences. **(b)** Log-log plot showing the relationship between coarticulatory magnitude and articulatory vowel-consonant (V-C) distance. The magnitude of vowel coarticulation (i.e., how much vowels are influenced by adjacent consonants) correlates with the articulatory V-C distance. The blue fit line represents articulatory coarticulatory magnitude, while the red fit line represents acoustic coarticulatory magnitude.

3.6 Discussion

We present a methodology for the co-analysis of the complex articulatory mechanics of coarticulation and their acoustic realizations. Our analyses show that geometric morphometrics (GM) coupled with multivariate acoustic analysis is well suited to study the intricate coarticulatory relationships between speech sounds. Notably, we apply MFCCs as a complementary and more comprehensive spectral measure compared to traditional formants, enabling analysis of both vowel and consonant acoustics. Our results corroborate early findings by Lindblom's (Lindblom, 1963) target-undershoot model and Öhman's (Öhman, 1967) prediction of articulatory and phonetic shift of phonemes towards their neighbors. However, rather than simply confirming these observations, our study also highlights how Lindblom's and Öhman's ideas converge into a broader hypothesis of network-like interdependencies

of phones. Specifically, Lindblom's emphasis on reduced vowel targets during continuous speech, and Öhman's focus on coarticulatory "pull" between phonemes, each illuminate different facets of a unified dynamic in which all phones (vowels and consonants alike) adjust their relative positions while preserving contrast. The extent to which consonants "move" around in articulatory and acoustic space has not yet been empirically studied because consonants are harder to analyze in articulatory and acoustic spaces (Figueroa and Kim, 2021). Our study offers a first step in this direction by modeling consonant distributions as geometric shapes in articulatory and MFCC spaces. We find that consonants form a coherent polygon (e.g., p–t–k–s–l), which – like the vowel triangle (i–a–u) – shows a geometric correspondence between shape space and acoustic space. To the best of our knowledge, consonant polygons have not previously been characterized in this way, and their apparent geometric consistency across spaces offers a previously unexplored characterization of consonant organization.

Our findings suggest that while coarticulation leads to a distortion of the vowel triangle (i-a-u) and to the distortion and transposition of the consonant polygon (p-t-k-s-l), the relative locations of vowels and consonants in respect to each other remain fairly constant – irrespective of their absolute locations in shape space and acoustic space. According to their articulatory context, it appears that these phone constellations (the VCV triplets studied here) are transposed in shape space and adjusted relative to each other. In functional terms, this implies that articulatory relationships (or contrasts) between phones in shape space, rather than absolute articulatory phoneme positions, are of prime importance during speaking. In continuous speech production, the articulatory and acoustic contrasts between phonemes are therefore more important than their absolute morphological and acoustic positions in their respective spaces. According to this view, contrasts would provide "sufficient discriminability", as proposed by Lindblom (1990a). Thus, the hypothesis of relative articulatory invariance proposed here is in agreement with the general observation that acoustic perception – like any other form of sensory perception – is based on relative rather than absolute signal properties (cf., McMurray, 2022).

Our analyses also reveal that vowel articulation and acoustics in general are more constrained as compared to the articulation of consonants. Although the vowel triangle is subject to extensive alterations, leading to consonant-context dependent within- and between vowel variation, the positions of the cardinal vowels in articulatory and acoustic space are fairly constant – whereas consonant subspaces appear to be rather free to drift around in articulatory and acoustic space, without affecting their quality in conveying their targeted phonetic information. Nonetheless, our results suggest that coarticulatory effects are a driver of acoustic diversity within speaker communities, which could in turn condition the diversification of languages (Ohala, 1989; Moran et al., 2024). The methodology and results presented here may therefore shed light on the processes of sound change and language change in light of coarticulation effects.

Although numerous case studies document coarticulation in specific contexts, overarching theories that integrate the wide range of observations remain limited (e.g., Saltzman and Munhall, 1989; Lindblom, 1990a; Browman and Goldstein, 1992). Our morpho-phonetic framework contributes to broader theoretical developments by treating phone sequences as trajectories in articulatory and acoustic spaces, modeled with geometric morphometrics and quantitative acoustic measures. This approach moves beyond the concept of isolated phonemes, providing a dynamical account of how shape and position of phones shift due to coarticulation. By uniting

articulatory goals (e.g., forming specific constrictions) with higher-level coordination principles, our method builds on gestural and task-dynamic insights to capture both local coarticulatory effects and broader trade-offs between articulatory ease and phonetic clarity, thus echoing Lindblom's emphasis on balance between clarity and economy in speech production and perception (Lindblom, 1990a).

Similar processes shape the acoustic diversification of birdsong and spoken language

4.1 Context

The following manuscript with the title *Similar processes shape the acoustic diversification of birdsong and spoken language* is currently in preparation for submission to a scientific journal. This study was conducted by the author in collaboration with Steven Moran, Marcia Ponce de León and Christoph Zollikofer.

4.2 Abstract

Birdsong and spoken language share uncanny similarities in their structure, development, vocal production, function, and cultural transmission (Doupe and Kuhl, 1999; Salwiczek and Wickler, 2004; Bolhuis et al., 2010; Bolles, 2011; Hyland Bruno et al., 2021; Zhang et al., 2023; Longtine et al., 2024; Adam et al., 2025). Both birdsong and language have discrete hierarchical properties, whose variation across groups is well studied (Gray and Atkinson, 2003; Potvin and Clegg, 2015; James et al., 2021). However, their acoustics show continuous variation within and between groups (Clopper et al., 2005; Rivera et al., 2023), and the factors driving this variation are less well understood. This raises the question of whether the observed patterns of variation primarily reflect discrete categorical differences or whether they emerge from more continuous, gradual evolutionary processes (Parravicini and Pievani, 2018). Here we analyze acoustic variation in chaffinch birdsong and human vowel systems across populations in Eurasia. By combining multivariate spectral analyses with methods from evolutionary population biology, we find consistent patterns of acoustic variation within and between populations. Our results suggest that neutral-like evolutionary processes and isolation by distance underlie acoustic diversification in both birdsong and spoken language, and contribute to the broader debate on whether language and birdsong evolution is best understood as a series of discrete shifts or as a continuous, gradual process influenced by population-level factors.

4.3 Introduction

Spoken language is comprised of consonants, vowels, and tones, which form the basic building blocks of syllables, words and phrases. Birdsong is also syntactically structured, composed of syllables, phrases, and motifs that are arranged – as in different languages – into distinctive patterns. These distinctive patterns reflect "dialectal" differences between populations and across geographical distances, which both speakers and songbirds recognize (Podos and Warren, 2007; Koile et al., 2022); birdsong at the species and subspecies levels (Il'ina et al., 2021) and speakers in terms of

regional accents that eventually become mutually unintelligible languages (Purnell et al., 1999).

The discrete categories and hierarchical properties of language and birdsong lend themselves well to phylogenetic analysis, resulting in "phyletic trees" of language history and birdsong evolution (Gray and Atkinson, 2003; James et al., 2021). However, due to the continuous nature of vocalization, the diversifying patterns of acoustic variation – within and between populations in both systems – cannot be readily studied using standard phylogenetic methods relying on discrete trait data. Thus, the processes generating acoustic diversity of languages and birdsong remain largely unknown. This lacuna is striking because vocal learning is a rare trait among non-human species, and studying parallels in the acoustic diversification of these complex systems may provide a broad view on how acoustic communication systems evolve.

Languages and birdsong are both culturally transmitted systems that diversify over time. To better understand this diversification, we apply concepts from evolutionary population biology to the cultural evolution of language and birdsong. In particular, we draw on models originally developed for the evolution of genetic traits to explore how cultural traits may change under neutral versus selective pressures. Evolutionary change – whether biological or cultural – relies on within-population variation as the raw material for population-level divergence (Lande, 1979). Under neutral processes such as genetic drift, we expect a proportional relationship between within- and between-population variation (Lande, 1980; Ackermann and Cheverud, 2004), and we often observe that differences accumulate with geographic distance, a pattern known as isolation by distance – a hallmark of neutral drift in spatially structured populations (Lande, 1991; Slatkin and Excoffier, 2012). When this proportionality breaks down, it suggests that additional factors such as adaptation to local environments or social structures may be influencing the trajectory of evolution. This framework allows us to test whether the diversification of cultural traits like language and birdsong can be explained by neutral processes alone, or whether adaptive pressures also play a role (Lynch and Baker, 1993; Pawlowitsch et al., 2011; Potvin and Clegg, 2015; Bentz et al., 2018).

This approach has recently been applied to spoken language, where acoustic features are transmitted culturally rather than genetically (Moran et al., 2024). Similarly, birdsong – also learned and culturally transmitted (Riebel et al., 2015) – can be analyzed through this lens. If birdsong evolves in a neutral-like way, we should expect the same proportionality between within-population and between-population variation in acoustic features (var_w and var_b respectively), as well as isolation by distance. Deviations from this expectation would therefore suggest non-neutral processes, such as environmental adaptation or cultural processes.

In this study, we test this hypothesis in both humans and birds. We focus on a single language family and a single bird species complex, drawn from broadly overlapping geographic regions. This allows for a direct comparison of neutral versus adaptive patterns in learned acoustic systems.

4.4 Results

Chaffinches are widespread passerine birds that form a species-subspecies complex within the *Fringilla* genus. Their songs are ideal for comparison with spoken languages because they are culturally transmitted, show extensive geographic variation, and have been widely studied in terms of diversity, acquisition, and evolution

(Slater et al., 1984; Lachlan, 2000; Lachlan and Servedio, 2004; Lachlan, 2008; Lachlan et al., 2013; Riebel et al., 2015). Speech likewise exhibits population-level differentiation, with culturally transmitted sound systems that evolve over time.

We analyzed individual songs recorded from eight chaffinch populations and individual speech sounds from 15 speech communities, sampled across similar geographic areas (Fig. 4.1A; see SI Table 4.S1 for sample structure). For chaffinches, we used the Fourier-transformed spectrum as acoustic markers (Fig. 4.1B). For spoken language, we analyzed formant data from the three cardinal vowels (i, a, u) (Fig. 4.1C) in Indo-European languages. We included Basque speakers as a non Indo-European language in this analysis because of widespread bilingualism, allowing us to examine potential cross-language influences.

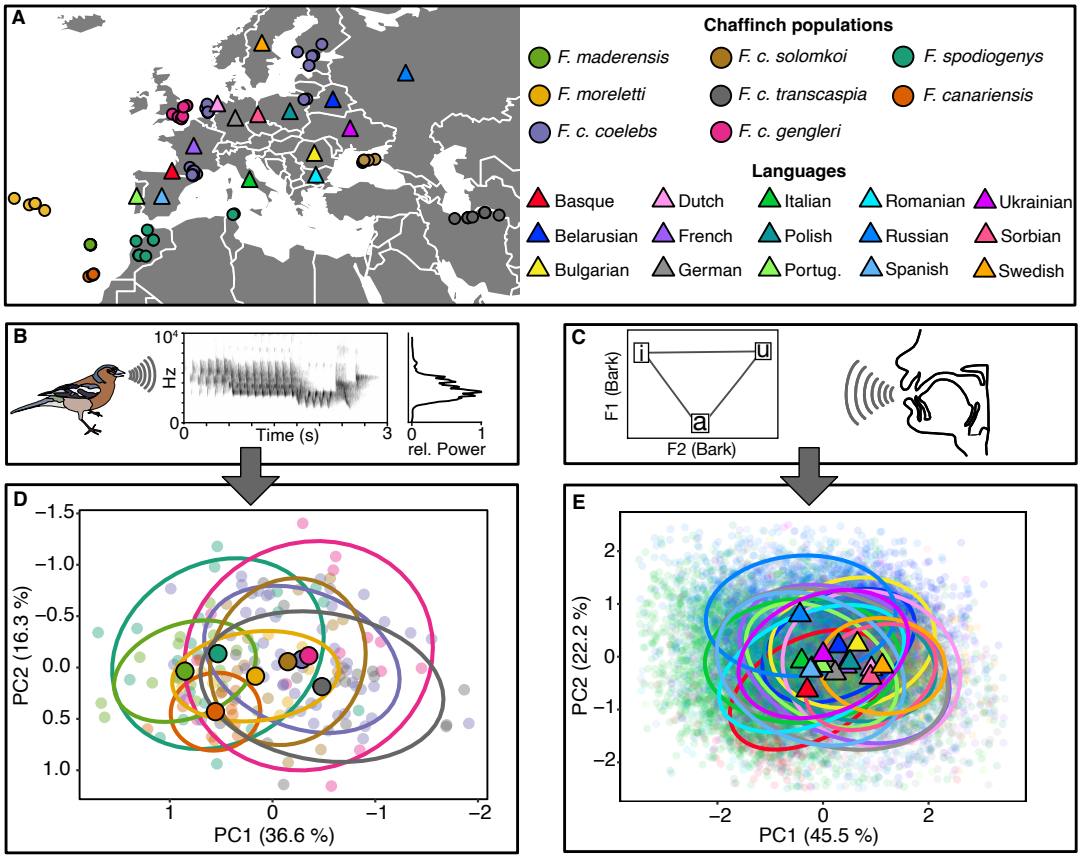


FIGURE 4.1: **Patterns of acoustic variation within and between populations of chaffinches and speakers. A**: Geographic distribution of acoustic samples of chaffinches and spoken languages **B**: For each individual birdsong sample, we calculated the average frequency spectrum of the time-domain acoustic signal, which captures the average power distribution across acoustic frequencies for each song (example spectrogram and frequency spectrum of *Fringilla coelebs* song). **C**: For each individual speech sample, we calculated the vowel configuration (i, a, u) in F1-F2 formant space. **D**–**E**: Principal component analysis (PCA) is used to visualize major patterns of variation within and between populations in the song spectra (**D**) and in the vowel configurations (**E**). Within-group variation is represented by 70% density ellipses drawn around group means.

We used PCA to visualize and explore the principal patterns of variation in the multivariate acoustic spaces of birdsong and vowels (Fig. 4.1D–E). We then analyzed intra- and interpopulation variation to assess consistency with expectations under neutral evolution. Finally, we tested for isolation by distance effects by evaluating whether acoustic dissimilarity increases with geographic separation. Our main workflow is summarized in Fig. 4.1 and described in detail in the *Materials and Methods* sections.

Fig. 4.1D shows principal patterns of variation in multivariate acoustic space for the different chaffinch populations. Each data point represents the specific acoustic makeup of an individual song. Populations differ from each other in their location in acoustic space, indicating different population-mean song spectra. African-Macronesian (*F. spodyogenys*, *F. canariensis*, *F. maderensis*, *F. moreletti*) and European populations (*F. c. gengleri*, *F. c. coelebs*, *F. c. solomkoi*, *F. c. transcaspia*) each form clusters that largely reflect geographic proximity. However, there is substantial overlap between populations, suggesting a continuum of song variation rather than strictly discrete song categories.

Fig. 4.1E shows principal patterns of variation in multivariate acoustic space for the different speech communities. The individual data points refer each to an individual speaker's vowel representation of the cardinal vowels in formant space (Fig. 4.1C). As in birdsong, there is substantial overlap in acoustic variation between languages (Fig. 4.1E), with partially overlapping distributions reflecting shared variability alongside differences in their average vowel configurations.

In both birdsong and speech data, acoustic variation is continuous across groups, with partial overlap in their multivariate distributions. Quantitatively, within-group variance ($var_{\text{song.w}}$, $var_{\text{lang.w}}$) far exceeds between-group variance ($var_{\text{song.b}}$, $var_{\text{lang.b}}$) (see SI Table 4.S2), consistent with a scenario in which most acoustic diversity emerges within populations rather than across them – a typical pattern expected for evolutionary processes, where within-group variation feeds between-group diversification.

Moreover, a comparison of within-population variation (var_w) and between population variation (var_b) in chaffinch populations and speech communities shows that patterns of variation within groups and between groups do not deviate from proportionality, i.e., the principal variation vectors of within-group and between-group variation are statistically indistinguishable from each other (Figs. 4.2A–B; multivariate comparison of variances: p_{song}=0.24 and p_{lang}=0.52; see also Table 4.1 and *Materials and Methods*). This finding suggests that the acoustic variation observed within groups mirrors the acoustic variation seen between groups, both in chaffinch song and spoken language.

To substantiate this observation, a regression analysis on $\log_{10}$-transformed within- and between-group variation in multivariate acoustic space shows a slope near 1 for both chaffinch populations and language communities (Table 4.1 & Fig. 4.S1). A slope close to 1 suggests a proportional relationship, indicating that the range of variation observed within groups predicts the degree of variation seen between groups (Lande, 1980; Ackermann and Cheverud, 2004). This pattern implies that the same processes generating diversity within groups also contribute to differences between groups, supporting a neutral-like mode of acoustic differentiation.

Furthermore, we find evidence of isolation by distance in both the chaffinch acoustic data and language acoustic data (Fig. 4.2C–D), where acoustic distance between populations increases gradually with their geographic distance. Acoustic differentiation between chaffinch populations is also stronger when comparing island

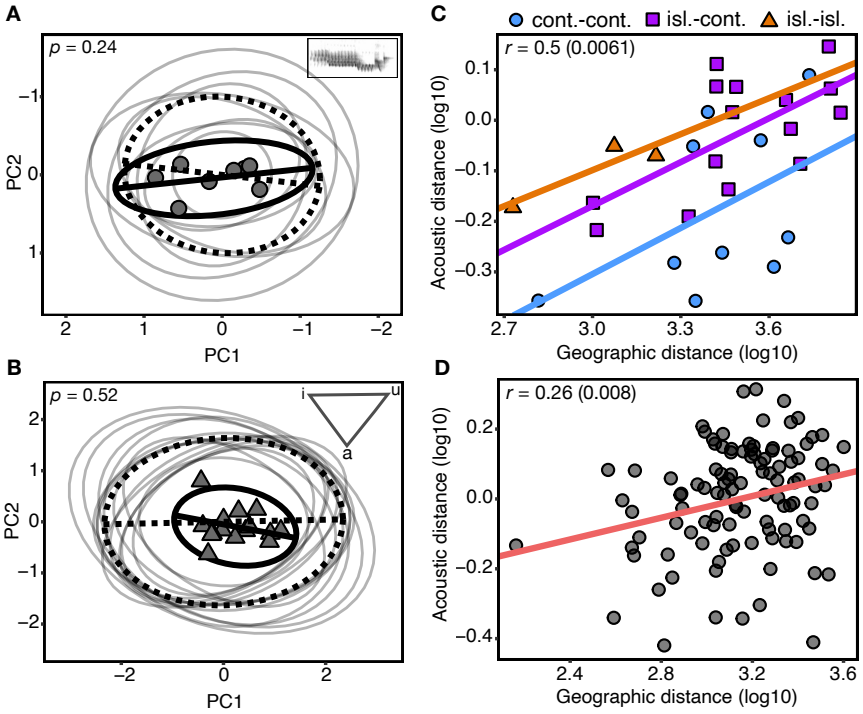


FIGURE 4.2: **Neutral evolution and isolation by distance. A–B**: Multivariate acoustic variation is partitioned into pooled within-population variation (dashed black ellipses) and between-population variation (solid black ellipses) for chaffinch song spectra (**A**) and speaker vowel spaces (**B**). The lines spanning each pair of ellipses represent the principal axes of variation: within-population (dashed line) and between-population (solid line). Statistical tests indicate that these axes do not significantly deviate from proportionality, consistent with expectations under neutral evolution (see *p*-values in panels **A**, **B**, and Table 4.1). Gray ellipses represent the centered 90% density ellipses for individual chaffinch and speaker populations. **C–D**: Isolation by distance patterns among chaffinch populations (**C**) and speaker populations (**D**). Acoustic distances are Euclidean distances between group means in multivariate acoustic space (see Figs. 4.1D, E). In the chaffinch data (**C**), pairwise comparisons are color-coded: island–island (orange triangles), island–continent (purple squares), and continent–continent (blue dots), with separate linear fits shown for each. The correlation coefficients (*r*) refer to the linear fits across all comparisons (associated *p*-value in brackets).

vs. continental populations and island vs. island populations, consistent with expectations under neutral processes. Since islands tend to be more geographically and ecologically isolated, such patterns are frequently observed in systems where limited gene flow or cultural exchange reinforces divergence (Newton, 2003; Bromham et al., 2024).

4.5 Discussion

In many areas of linguistic research, including dialectology, sociolinguistics, and automatic speech recognition, variation between speaker groups is often the primary research focus. Hence, variation within groups is frequently treated as noise, as it tends to obscure differences between groups (Labov, 1966; Lobanov, 1971; Adank et al., 2004; Giuliani et al., 2006; Burridge, 2023).

In contrast, the approach taken here treats within-group acoustic-phonetic variation as data rather than noise, and as a potential source of between-group variation. Our results suggest that both birdsong and spoken language undergo neutral-like

Chaffinch population	p	Slope	**Language**	p	Slope
F. spodiogenys	0.09	1.24 (0.46)	*Basque*	0.07	0.83 (0.47)
F. canariensis	0.17	1.91 (0.33)	*Belarusian*	0.09	0.90 (0.55)
F. c. coelebs	0.21	1.76 (0.04)	*Bulgarian*	0.77	0.84 (0.40)
F. c. gengleri	0.17	1.12 (0.77)	*Dutch*	0.50	1.34 (0.58)
F. maderensis	0.02	1.37 (0.43)	*French*	0.66	1.91 (0.23)
F. moreletti	0.14	1.51 (0.19)	*German*	0.79	1.18 (0.71)
F. c. solomkoi	0.02	1.19 (0.69)	*Italian*	0.73	0.93 (0.83)
F. c. transcaspia	0.18	1.17 (0.59)	*Polish*	0.94	1.09 (0.85)
			Portuguese	0.40	1.87 (0.12)
			Romanian	0.08	0.93 (0.84)
			Russian	0.70	0.98 (0.93)
			Sorbian	0.34	0.66 (0.38)
			Spanish	0.56	1.39 (0.51)
			Swedish	0.70	1.35 (0.20)
			Ukrainian	0.92	1.15 (0.77)

TABLE 4.1: **Comparison of variance-covariance patterns between intra- and inter-group data for chaffinch populations and language data**. *P* represents the *p*-value from the eigenanalysis test, i.e., a multivariate comparison of variances (Le Maître and Mitteroecker, 2019). The slopes are the empirically derived values from regressions of between-population variation on within-population variation, with values in parentheses indicating *p*-values for tests of deviation from a slope of 1 (see Materials and Methods; *Testing evolutionary hypotheses* for details and SI Fig. 4.S1).

processes of acoustic diversification, where variation between populations largely arises from the internal diversity within each group and in which random processes, rather than selective pressures, drive the divergence in acoustic traits across populations. These findings contribute to the broader debate on whether language and birdsong evolution is best understood as a series of discrete shifts or as a continuous, gradual process influenced by population-level factors (Parravicini and Pievani, 2018). By showing that within-group variation conditions between-group divergence, our results suggest that the evolution of sounds and songs is shaped by cumulative, stochastic processes rather than abrupt categorical changes.

Thus, we suggest that neutral processes drive gradual acoustic divergence over time in chaffinch populations and speech communities. This neutral framework, together with isolation by distance, provides a baseline for understanding acoustic differentiation by enabling the evaluation of alternative hypotheses, such as the role of stronger environmental or cultural adaptations in shaping observed patterns. These effects are testable within our null model of neutral-like evolution, where deviations from the neutral model may reveal their role in shaping acoustic differences in both spoken language and birdsong (Payne, 1978; Krause, 1987; Rothstein and Fleischer, 1987; Krause, 1993; Nelson and Marler, 1994; Podos and Warren, 2007; Benítez-Burraco and Moran, 2018; Benítez-Burraco and Moran, 2024).

For instance, studies show that acoustic signals adapt to ecological conditions, a phenomenon known as the Acoustic Adaptation Hypothesis (AAH). AAH has been empirically studied across various animal groups, especially birds (Chapuis, 1971; Morton, 1975; Seddon, 2005; Boncoraglio and Saino, 2007), where specific spectral and temporal adjustments of birdsong correspond with habitat typesNottebohm, 1969; Boncoraglio and Saino, 2007. In particular, birds in dense vegetation or urban

settings adapt their songs to compensate for environmental noise and habitat density (Moseley et al., 2018; De Framond and Brumm, 2022). This observation aligns with work showing that languages in diverse physical environments may exhibit different phonological patterns – suggesting adaptations to external conditions. For example, by quantifying whether languages spoken in environments with denser vegetation and higher ambient temperatures use more sonorous sounds, given that these factors impede the transmission of higher frequencies (Maddieson and Coupé, 2015). This recent renaissance in asking why linguistic diversity is distributed unevenly across the globe has been in turn due to increased open access to large linguistic databases (Moran and McCloy, 2019; Hammarström et al., 2024), allowing researchers to investigate which non-linguistic factors, such as genes, culture, environment, climate, population movements, and demography among others, correlate with patterns of phonetic and phonological diversity (Pericliev, 2004; Hay and Bauer, 2007; Donohue and Nichols, 2011; Atkinson, 2011; Moran et al., 2012; Everett, 2013; Greenhill, 2014; Everett et al., 2015; Creanza et al., 2015; Maddieson, 2018; Blasi et al., 2019; Urban and Moran, 2021; Maddieson and Benedict, 2023; Liang et al., 2023).

Lastly, our research broadens our understanding of how acoustic signals evolve and diversify in species that are vocal learners (Lynch and Baker, 1993; Lynch, 1996; Salwiczek and Wickler, 2004; Podos and Warren, 2007; Bolhuis et al., 2010), supporting the view that biological, cultural and environmental factors are deeply interwoven in shaping complex vocal communication systems. Notably, the Basque speakers in our sample cluster closest to speakers of Spanish, Romanian, Portuguese, and Italian in multivariate acoustic space (Fig. 4.1E), despite the distinct linguistic lineage of Basque. This pattern is consistent with extensive bilingualism and potential cross-linguistic influence on vowel production. Future research should explore these interactions further, specifically examining the roles of contact situations as well as genetic, morphological, and environmental constraints alongside neutral processes across vocal learning species. This balanced approach may illuminate how random drift, isolation by distance, and adaptation together shape the complex landscape of vocal communication across taxa, perhaps offering a unified perspective on acoustic diversification.

4.6 Materials and Methods

Birdsong data

The birdsong data sample consists of N=163 individual recordings from xeno-canto.org (XC), comprising different species and subspecies within the chaffinch complex (genus *Fringilla*; see Fig. 4.1A and Table 4.S1). The groups are *Fringilla spodiogenys* (Morocco and Tunisia), *Fringilla canariensis* (Canaries), *Fringilla maderensis* (Madeira), *Fringilla moreletti* (Azores), *Fringilla coelebs coelebs* (nominate; N-EU, Netherlands, France), *Fringilla coelebs gengleri* (UK), *Fringilla coelebs solomkoi* (Crimea) and *Fringilla coelebs transcaspia* (Iran). Note that the Macaronesian populations (*F. canariensis*, *F. maderensis*, *F. moreletti*) have only recently been suggested to represent separate species (Recuerda et al., 2021).

Each birdsong recording was visualized as a spectrogram (Fig. 4.1B) and subsequently extracted as a separate sound file in .wav format using the software Praat (Boersma, 2001). Only song sequences with decent quality and minimal background noise were considered. Song sequences with aberrant song were excluded, as well as recordings with non-chaffinch song or other song/calls in the background. We

applied noise reduction to each song sequence and used high-pass filtering in Audacity to remove remaining background noise from the samples. The source recordings with the XC data ID, audio recordist name and direct URL to each recording is given in Supplementary Table 4.S3.

Spoken language data

The vowel dataset is taken from the VoxCommunis corpus (Ahn and Chodroff, 2022), which contains phone-level alignments and acoustic-phonetic measurements derived from Mozilla Common Voice Corpus (Ardila et al., 2020). We calculated the vowel space from five cardinal vowels (i, a, u) for each individual in the sample for 15 languages (Basque, Belarusian, Bulgarian, Dutch, French, German, Italian, Polish, Portuguese, Romanian, Russian, Sorbian, Spanish, Swedish, Ukrainian; see SI Table 4.S1) for subsequent multivariate acoustic analysis (see Fig. 4.1). We included only speakers that produced the complete set of vowels, which resulted in a sample of N=55,116 data points (vowel utterances) from N=18,372 individuals, i.e., their individual vowel spaces.

Analysis and visualization of birdsong data

We used the R package *seewave* (R Core Team, 2024; J. Sueur et al., 2008) to extract the mean spectrum along each recording in the sample (N=163), which returns the mean relative amplitudes of the frequency distribution over the complete song via the Fourier transform. Each individual mean spectrum was normalized by its maximum in order to account for amplitude differences across individual songs. For the calculation of the mean spectra, a number of K=256 frequency bins were chosen. The resulting data matrix (i.e., the mean spectra) therefore consisted of N=163 rows (representing the N=163 individual recordings) and K=256 columns (representing the K=256 frequency bins). This N x K matrix was subjected to principal component analysis (PCA) in order to generate a low-dimensional representation of multivariate acoustic space (Fig. 4.1D). The total variation in the multivariate acoustic space was then subdivided into pooled within-population variation ($var_{\text{song.w}}$) that is common to all populations, and between-population variation ($var_{\text{song.b}}$), as shown in Fig. 4.2A. To compute $var_{\text{song.w}}$, the PCA data was pooled across all populations (N=163), whereas $var_{\text{song.b}}$ was assessed by leveraging the mean location in multivariate acoustic space for each population (N=8).

Analysis and visualization of spoken language data

The formant data was Bark-scaled because human hearing is logarithmic in terms of intensity (loudness) and frequency (Traunmüller, 1990). To compare the within and between-language variation in the sample, the vowel space data was expressed as data vectors of length $3Q$ (3 vowels, Q=2 formants F1-F2). PCA was then applied to the resulting 18,372 x 6 data matrix (N=18,372 individuals x $3Q$) to provide a low-dimensional representation of multivariate acoustic space (Fig. 4.1E). The total variation in the PCA data was subdivided into within-language variation ($V_{\text{lang.w}}$; N=18,372), shared by all languages, and between-language variation ($V_{\text{lang.b}}$, N=15), using a similar approach as for the chaffinch song data (Fig. 4.2B).

Testing evolutionary hypotheses

We conducted a comparison of the primary patterns of acoustic variation within the chaffinch song multivariate space and the vowel multivariate acoustic space (Fig. 4.2), examining both within-population ($var_{\text{song.w}}$ and $var_{\text{lang.w}}$) and between-population ($var_{\text{song.b}}$ and $var_{\text{lang.b}}$) variation. We overlaid the between-group and pooled within-group multivariate acoustic spaces (see Figs. 4.2A–B). The principal variation vectors, which represent the major axes of the density normal ellipses for the within-group and between-group variation, exhibit nearly collinear alignment, suggesting that the variance-covariance matrices of these distributions are nearly proportional. We statistically confirmed this proportionality using eigenanalysis-based tests with the R package *vcvcomp* (Bookstein and Mitteroecker, 2014; Le Maître and Mitteroecker, 2019). Detailed statistics (p-values) for each population/speech community are provided in Table 4.1.

To compare the magnitudes of within-group and between-group variance (Table 4.S2), we computed the aggregated variances over all principal components in the chaffinch multivariate acoustic space and vowel acoustic multivariate space (Figs. 4.1D & 4.1E). Between-group variance was computed using the population-means in the respective multivariate acoustic space.

Following the analogy of neutral phenotypic evolution models (Lande, 1980; Ackermann and Cheverud, 2004), we hypothesized that under neutral-like conditions of birdsong and spoken language acoustic diversification, the pattern of acoustic variation between groups ($var_{\text{song.b}}$, $var_{\text{lang.b}}$) should be proportional to the variation within groups ($var_{\text{song.w}}$, $var_{\text{lang.b}}$) (Lande, 1980; Ackermann and Cheverud, 2004; Moran et al., 2024). This hypothesis was assessed by orthogonalizing the variance-covariance matrices, characterizing each between-group and within-group variation, identifying statistically independent variance components along their eigenvectors. Subsequently, we evaluated the proportionality between these matrices by regressing the between-group variance components against the within-group components. Proportionality is indicated by linearity (approximating a slope of 1 in a double-logarithmic plot (Fig. 4.S1). Note that we selected only the first 10 PCs of the birdsong data for this analysis, which together comprise over 90% of the total variation, to minimize the influence of higher-order PCs that primarily reflect noise rather than meaningful spectral information (variance proportions comprised in PCs are shown in Fig. 4.S2). Regression statistics for each population are provided in Table 4.1.

To calculate the correlation between the acoustic distances and geographic distances between groups (chaffinch populations and speech-communities, respectively; see Fig. 4.2C–D), i.e., to assess isolation by distance, we computed the Euclidean distances between the group-means for each group-pair in multivariate acoustic space (Fig. 4.1D–E). The geographic distances were approximated by calculating the geodesic distances between each group-centroid (population/language) (Hijmans, 2024), i.e., the centroid-location of the chaffinch populations and the centroid-location of the sampled speech-communities.

4.7 Supplementary Information

Supplementary Figures

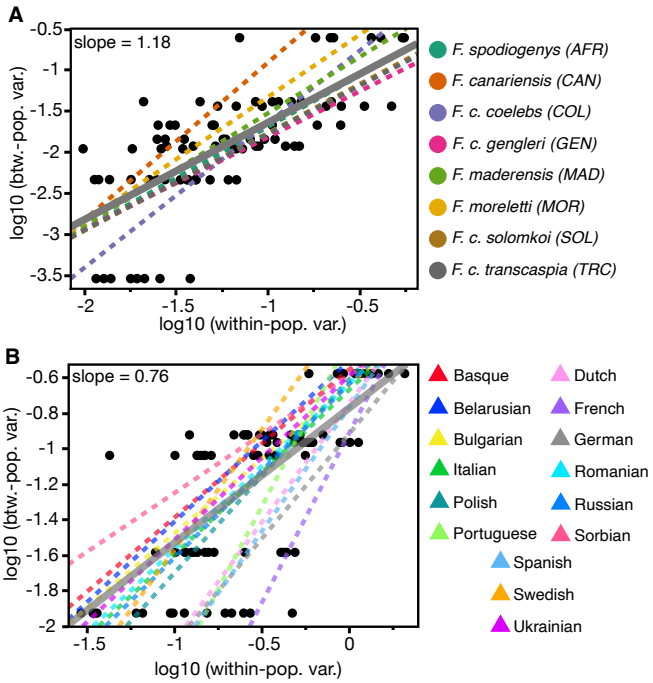


FIGURE 4.S1: **Linear regression of between-group variance on within-group variance for the chaffinch data (A) and the language data (B)**. The gray regression lines indicate the linear fit for the pooled data, the dashed lines (colored) represent each an individual population (see legends on the right). Slopes and p-values are provided in Table 4.1.

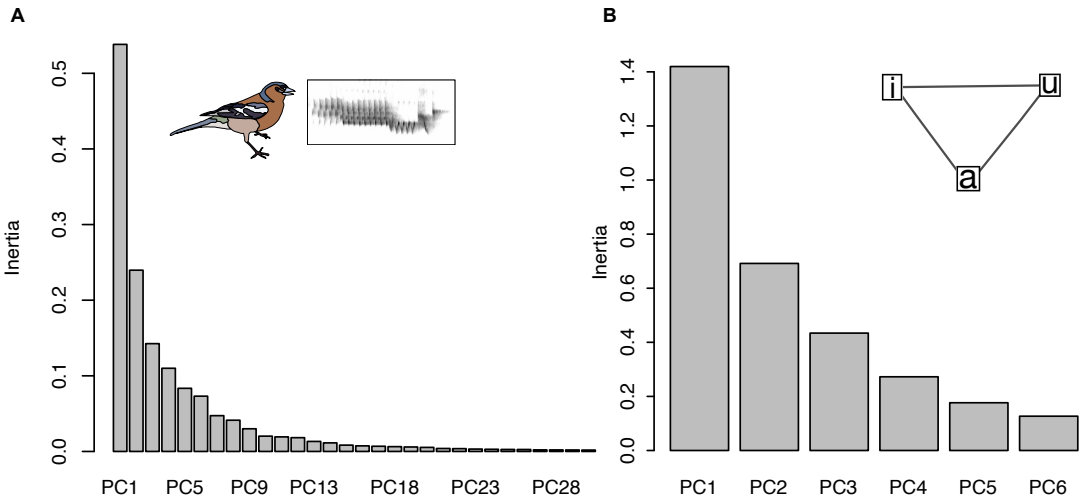


FIGURE 4.S2: Variance proportions (interia) comprised by the PCs of the spectral birdsong analysis (**A**) and the vowel space analysis (**B**). Note that for birdsong (**A**) we only show the first 30 PCs.

Supplementary Tables

Chaffinch population	*N*	**Language**	*N*
F. spodiogenys	21	*Basque*	812
F. canariensis	12	*Belarusian*	3610
F. c. coelebs	56	*Bulgarian*	35
F. c. gengleri	15	*Dutch*	105
F. maderensis	16	*French*	654
F. moreletti	8	*German*	538
F. c. solomkoi	16	*Italian*	6033
F. c. transcaspia	19	*Polish*	2523
		Portuguese	392
		Romanian	190
		Russian	1463
		Sorbian	18
		Spanish	837
		Swedish	592
		Ukrainian	570
Total	163	Total	18372

TABLE 4.S1: Overview of the sample structure for chaffinch populations (left) and speech communities (right), detailing the number of individuals or speakers sampled per group. Notably, sample size across groups does not significantly correlate with within-population variance for either the chaffinch or language data, indicating that the observed acoustic diversity within populations is not a function of sample size but represents inherent variation.

Chaffinch population	$var_{\text{song.w}}$	$var_{\text{song.w}}/var_{\text{song.b}}$
F. spodiogenys	1.36	3.33
F. canariensis	0.47	1.15
F. c. coelebs	1.20	2.95
F. c. gengleri	1.67	4.10
F. maderensis	0.89	2.19
F. moreletti	0.78	1.92
F. c. solomkoi	1.30	3.18
F. c. transcaspia	1.57	3.84
$var_{\text{song.b}}$	0.408	
Population-mean	1.16	2.83

Language	$var_{\text{lang.w}}$	$var_{\text{lang.w}}/var_{\text{lang.b}}$
Basque	1.94	3.13
Belarusian	1.87	3.02
Bulgarian	2.49	4.02
Dutch	3.47	5.62
French	4.96	8.03
German	4.83	7.82
Italian	2.78	4.49
Polish	2.50	4.05
Portuguese	2.63	4.26
Romanian	2.51	4.05
Russian	2.70	4.37
Sorbian	1.34	2.16
Spanish	3.72	6.02
Swedish	1.53	2.48
Ukrainian	2.29	3.71
$var_{\text{lang.b}}$	0.62	
Language-mean	2.77	4.48

TABLE 4.S2: Comparison of within-group (var_{w}) and between-group (var_{b}) variance in the sampled chaffinch populations (left) and speech communities (right). For each population, the calculated within-group variances and the ratio of within-group vs. between-group variance is given. The second to last row ($var_{\text{song.b}}$; $var_{\text{lang.b}}$) shows the between-group variance. The last row refers to the mean within-group variance and mean variance-ratio across all groups.

ID	Recordist	URL
XC167570	Max Allan Niklasson	https://xeno-canto.org/167570
XC168828	Johannes Honold	https://xeno-canto.org/168828
XC181893	Thijs Fijen	https://xeno-canto.org/181893
XC181919	Thijs Fijen	https://xeno-canto.org/181919
XC241962	Terje Kolaas	https://xeno-canto.org/241962
XC314209	Alain Malengreau	https://xeno-canto.org/314209
XC410113	Cherdoud Ayoub	https://xeno-canto.org/410113
XC709520	Paweł Malczyk	https://xeno-canto.org/709520
XC834601	Helmut Schaffer	https://xeno-canto.org/834601
XC90859	Lars Lachmann	https://xeno-canto.org/90859
XC367194	Yoann Blanchon	https://xeno-canto.org/367194
XC45368	Christoph Bock	https://xeno-canto.org/45368
XC460042	Alain Verneau	https://xeno-canto.org/460042
XC472080	Rubén Barone	https://xeno-canto.org/472080
XC472082	Rubén Barone	https://xeno-canto.org/472082
XC483980	Rubén Barone	https://xeno-canto.org/483980
XC28239	Stuart Fisher	https://xeno-canto.org/28239
XC30118	Ruud van Beusekom	https://xeno-canto.org/30118
XC543792	Albert Noorlander	https://xeno-canto.org/543792
XC554923	Albert Noorlander	https://xeno-canto.org/554923
XC581763	Albert Noorlander	https://xeno-canto.org/581763
XC648240	Arend Wassink	https://xeno-canto.org/648240
XC654715	Albert Noorlander	https://xeno-canto.org/654715
XC710831	Albert Noorlander	https://xeno-canto.org/710831
XC164002	Tero Linjama	https://xeno-canto.org/164002
XC236816	Eetu Paljakka	https://xeno-canto.org/236816
XC237329	Eetu Paljakka	https://xeno-canto.org/237329
XC268097	Eetu Paljakka	https://xeno-canto.org/268097
XC349842	Tero Linjama	https://xeno-canto.org/349842
XC476674	Uku Paal	https://xeno-canto.org/476674
XC512404	Jarek Matusiak	https://xeno-canto.org/512404
XC793409	Jarek Matusiak	https://xeno-canto.org/793409
XC385395	Cedric Mroczko	https://xeno-canto.org/385395
XC396978	Cedric Mroczko	https://xeno-canto.org/396978
XC562081	Cedric Mroczko	https://xeno-canto.org/562081
XC576473	Cedric Mroczko	https://xeno-canto.org/576473
XC579065	Cedric Mroczko	https://xeno-canto.org/579065
XC579315	Cedric Mroczko	https://xeno-canto.org/579315
XC579323	Cedric Mroczko	https://xeno-canto.org/579323
XC579324	Cedric Mroczko	https://xeno-canto.org/579324
XC647075	Cedric Mroczko	https://xeno-canto.org/647075
XC653390	Cedric Mroczko	https://xeno-canto.org/653390
XC664513	Cedric Mroczko	https://xeno-canto.org/664513
XC789403	Cedric Mroczko	https://xeno-canto.org/789403
XC792675	Cedric Mroczko	https://xeno-canto.org/792675
XC177715	Dave Curtis	https://xeno-canto.org/177715
XC177716	Dave Curtis	https://xeno-canto.org/177716
XC256904	Bram Piot	https://xeno-canto.org/256904
XC29753	Stuart Fisher	https://xeno-canto.org/29753

Table 4.S3 continued from previous page

XC29754	Stuart Fisher	https://xeno-canto.org/29754
XC324043	Frank Lambert	https://xeno-canto.org/324043
XC324047	Frank Lambert	https://xeno-canto.org/324047
XC414237	Frank Lambert	https://xeno-canto.org/414237
XC894027	Stuart Fisher	https://xeno-canto.org/894027
XC189890	Lars Lachmann	https://xeno-canto.org/189890
XC315379	Cedric Mroczko	https://xeno-canto.org/315379
XC315381	Cedric Mroczko	https://xeno-canto.org/315381
XC316390	Cedric Mroczko	https://xeno-canto.org/316390
XC316391	Cedric Mroczko	https://xeno-canto.org/316391
XC316396	Cedric Mroczko	https://xeno-canto.org/316396
XC316400	Cedric Mroczko	https://xeno-canto.org/316400
XC316405	Cedric Mroczko	https://xeno-canto.org/316405
XC316415	Cedric Mroczko	https://xeno-canto.org/316415
XC794199	Carlos Pereira	https://xeno-canto.org/794199
XC798745	SonoNatura	https://xeno-canto.org/798745
XC805831	Carlos Pereira	https://xeno-canto.org/805831
XC812331	Carlos Pereira	https://xeno-canto.org/812331
XC812335	Carlos Pereira	https://xeno-canto.org/812335
XC812338	Carlos Pereira	https://xeno-canto.org/812338
XC109840	Eugenia Yablonovska-Grishchenko	https://xeno-canto.org/109840
XC109846	Eugenia Yablonovska-Grishchenko	https://xeno-canto.org/109846
XC109848	Eugenia Yablonovska-Grishchenko	https://xeno-canto.org/109848
XC413410	Kucherenko Volodymyr	https://xeno-canto.org/413410
XC418359	Kucherenko Volodymyr	https://xeno-canto.org/418359
XC463626	Kucherenko Volodymyr	https://xeno-canto.org/463626
XC468155	Kucherenko Volodymyr	https://xeno-canto.org/468155
XC479851	Kucherenko Volodymyr	https://xeno-canto.org/479851
XC243951	Cedric Mroczko	https://xeno-canto.org/243951
XC244387	Cedric Mroczko	https://xeno-canto.org/244387
XC474314	Ding Li Yong	https://xeno-canto.org/474314
XC654803	Ding Li Yong	https://xeno-canto.org/654803
XC727546	Cedric Mroczko	https://xeno-canto.org/727546
XC727693	Cedric Mroczko	https://xeno-canto.org/727693
XC731101	Cedric Mroczko	https://xeno-canto.org/731101
XC731331	Cedric Mroczko	https://xeno-canto.org/731331
XC734900	Ding Li Yong	https://xeno-canto.org/734900
XC743453	Cedric Mroczko	https://xeno-canto.org/743453
XC743607	Cedric Mroczko	https://xeno-canto.org/743607
XC745685	Cedric Mroczko	https://xeno-canto.org/745685

TABLE 4.S3: File ID, recordist and the URL of the chaffinch song recordings used in this study.

Gradual acoustic divergence of languages suggests acoustic data retains deep linguistic signals

5.1 Context

The following manuscript with the title *Evidence for gradual acoustic divergence suggests phonetic data retains deep linguistic signals* is currently in preparation for submission to a scientific journal. This study was conducted by the author in collaboration with Steven Moran, Marcia Ponce de León and Christoph Zollikofer.

5.2 Abstract

Linguistic diversity emerges from an interplay of biological, cultural, and environmental factors (Blasi et al., 2019; Dediu et al., 2019; Gisladottir et al., 2023; Liang et al., 2023), which shape the phonological repertoires, lexicons, and grammars of the world's languages as they evolve across space and time (Nichols, 1992). While traditional approaches in diachronic linguistics operate on discrete units, such as phonemes or lexical cognates (Atkinson and Gray, 2005; Greenhill et al., 2017; Barbieri et al., 2022), here we study synchronic acoustic variation in spoken languages, namely vowel systems. We analyze speech data from 18,372 individuals across 15 Eurasian languages (Indo-European and Basque) and apply methods from population biology to quantify within- and between-language acoustic diversity. Our findings reveal an isolation by distance (IBD) pattern in acoustic variation, indicating a process of gradual divergence shaped by drift-like dynamics. In contrast, lexical divergence does not consistently follow an isolation by distance pattern, perhaps due to the faster pace and stochastic nature of lexical turnover, which can obscure signals of shared history over time (Gray, 2005; Greenhill et al., 2017). Furthermore, similarity-based clustering of population-level vowel acoustics broadly recovers known genealogical relationships between language groups, implying that gradual acoustic divergence may track deep-time evolutionary relationships between languages and language groups. These results suggest that acoustic features of languages may preserve historical signals over greater time depths than lexical data, which is constrained by the time limitations of the comparative method (8–10k years) due to the compounding effects of sound change (Gray, 2005).

5.3 Introduction

Studies in diachronic linguistics typically use as input discrete properties of languages, such as phonemes or lexical cognates, to reconstruct the vocabulary and culture of ancient languages, or to identify the genealogical relatedness between languages to reconstruct language family phylogenies (Atkinson and Gray, 2005; Barbieri et al., 2022). Some studies have analyzed count-based measures, such as

"phonemic diversity", to assess global patterns of linguistic and genetic variation or serial founder effects and human dispersal out of Africa (Atkinson, 2011; Creanza et al., 2015). However, both approaches have limitations beyond the well known time-depth barrier of the historical comparative method (8-10k years), including: biases in phonemic analyses (Cysouw et al., 2012; Anderson et al., 2023); discrepancies in expert and computer-assisted cognate judgments (Rama et al., 2018); and contact-induced language change, which notoriously obscures population dynamics resulting in, for example, considerable mismatches between genetic and linguistic histories (Hunley et al., 2008; Pakendorf, 2014; Barbieri et al., 2022).

Although there are striking parallels between linguistic and biological evolution (Pagel, 2017; Ladoukakis et al., 2022), comparative linguistic approaches ignore the role of variation between individuals and speech communities in shaping language change (Fox, 1995; Campbell and Poser, 2008; Campbell, 2013). However, studying inter- and intra-group variation is essential for understanding evolutionary dynamics (Lande, 1980; Lande and Arnold, 1983; Lande, 1991; Ackermann and Cheverud, 2004; Ponce de León et al., 2018), because evolutionary processes feed on the variation within groups for between-group divergence (Ohala, 1989; Mesoudi et al., 2006; MacKenzie, 2019; Anthonissen, 2021; Yu, 2021). Recent research suggests that analyzing acoustic phonetic variation within and between speech communities provides new insights into spoken language diversification (Moran et al., 2024).

Here, we analyze vowel acoustics from N=18,372 speakers from 15 Eurasian languages (Indo-European and Basque; see Fig. 5.1A), using data from the Mozilla Common Voice corpus and extracted via the VoxCommunis pipeline (Ardila et al., 2020; Ahn and Chodroff, 2022). We focus on the three cardinal vowels (i, a, u) because they are nearly universal in the world's languages, which provides a baseline for cross-linguistic comparison, and they encode both speaker-specific and language-specific phonetic variation. For each speaker, vowel formant values (F1 & F2) are measured in the perceptual Bark scale and are used to construct individual acoustic profiles (Fig. 5.1B) (Traunmüller, 1990). These three-vowel acoustic profiles are then studied via principal component analysis (PCA) to identify the major patterns of variation across speakers and speaker groups in a low dimensional multivariate acoustic space, which allows us to compute population-level means and acoustic distances between languages (Figs. 5.1C–D; details in *Materials and Methods*).

This dataset allows us to test whether cross-linguistic patterns of acoustic variation reflect gradual diversification and spatial diffusion, as predicted by models of neutral evolution. Specifically, we examine whether acoustic similarity between languages results from neutral, drift-like processes, i.e., non-directional, cumulative changes over time akin to genetic drift in biology (Lande, 1991; Slatkin and Excoffier, 2012). A key expectation of such processes is isolation by distance (IBD): the principle that trait similarity decreases with geographic distance, as small, undirected changes accumulate and spread locally over generations. We test this by evaluating whether acoustic distance increases with geographic distance across a diverse sample of Indo-European languages, along with the Basque isolate. Basque serves as a useful test case because, while lexically unrelated to Indo-European languages (Trask, 2013), it has long been in contact with Spanish through widespread bilingualism – providing insight into how contact may locally shape acoustic similarity. Finally, we compare acoustic and lexical distances across the same languages to explore whether different components of language retain distinct traces of population history and linguistic evolution.

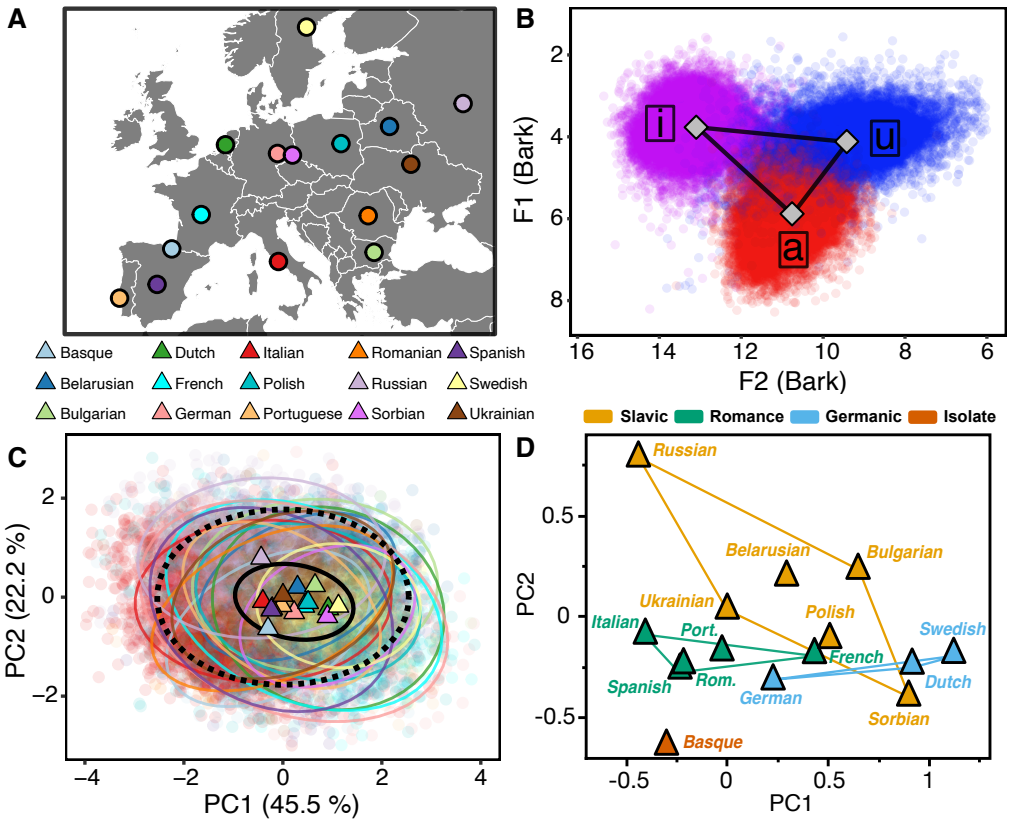


FIGURE 5.1: **Population-level analysis of acoustic data**. **A**: Sampled language communities and their location. **B**: The i-a-u vowel utterances (N=55,116) are shown in F1-F2 Bark scaled space. Diamonds indicate mean vowel positions, forming a vowel triangle. For each individual (N=18,372), we constructed a vowel triangle for principal component analysis (PCA; see **C**). **C**: PCA of individual vowel triangles, displayed with principal components PC1 and PC2. Each point in acoustic space represents an individual vowel space, while each colored triangle represents a language community centroid, i.e., the average vowel triangle per population. The 90% density ellipses indicate the major directions of variation within groups (pooled across languages; dashed black) and between groups (solid black). **D**: Acoustic variation of language-means (compare Fig. 5.1C), colored according to their genus within the Indo-European languages.

5.3.1 Acoustic variation reflects gradual diversification and isolation by distance

To understand how vowel acoustics diversify across languages, we first visualized the full dataset in multivariate acoustic space. The first two principal components (PCs 1 and 2) of multivariate acoustic space, which together explain 67.7% of the total variation, reveal a continuous, non-clustered distribution across speech communities (Fig. 5.1C). Rather than forming discrete groupings, acoustic variation across languages is gradual, suggesting gradual divergence. Notably, variation within speaker groups exceeds differences between them (Fig. 5.1C, Table 5.1). The within-group variation exhibits an approximately isotropic distribution, indicating that it is spread relatively evenly across dimensions of the acoustic space. Moreover, the within-group variation is proportional to the observed between-group differences. This relationship is visually apparent in the comparison of pooled within-group and

Language	N_A	$Var_{w.lang}$	Variance ratio
Basque	812	1.936693	3.132511
Belarusian	3610	1.873275	3.029937
Bulgarian	35	2.485738	4.020568
Dutch	105	3.472188	5.616104
French	654	4.965523	8.031505
German	538	4.83309	7.817302
Italian	6033	2.776998	4.491667
Polish	2523	2.503635	4.049515
Portuguese	392	2.634632	4.261397
Romanian	190	2.505825	4.053057
Russian	1463	2.703054	4.372066
Sorbian	18	1.335299	2.159786
Spanish	837	3.72067	6.018013
Swedish	592	1.530333	2.475244
Ukrainian	570	2.293467	3.709577

TABLE 5.1: **Sample sizes and variance metrics for acoustic data.** The table presents the sample sizes for acoustic data (N_A) across the studied speaker communities. $Var_{w.lang}$ represents the aggregated within-group variance of the acoustic data. The variance ratio (within-group variance relative to between-group variance) is provided for each language in the sample. Between-group variance was computed based on the mean values for each language. Notably, in all cases, within-group variance exceeds between-group variance, and there is no significant correlation between sample size and the magnitude of within-population acoustic variation (*see Materials & Methods*).

between-group ellipses shown in Fig. 5.1C. Such a pattern is consistent with theoretical expectations under a neutral model of phenotypic evolution, where diversification is primarily driven by stochastic, drift-like processes (Lande, 1980; Ackermann and Cheverud, 2004; Moran et al., 2024).

Language means in acoustic space broadly group by genus within the Indo-European family (Fig. 5.1D), though some languages (e.g., French, German, Sorbian) show shifts toward geographic neighbors. Furthermore, geographically close populations tend to be more acoustically similar (Figs. 5.1C–D), consistent with an isolation by distance signature. We find that acoustic variation consistently reflects isolation by distance: geographically closer languages tend to have more similar vowel acoustics (Fig. 5.2A). In contrast, lexical similarity shows this pattern only within closely related language genera, i.e., the isolation by distance signature disappears across deeper phylogenetic and time scales (Fig. 5.2B).

5.3.2 Acoustic similarity reflects genealogical relationships and language contact

To assess to what extent acoustic and lexical similarities reflect genealogical relatedness, we computed neighbor-nets based on between-language distances (Fig. 5.3). Note that, unlike traditional tree-based methods, the *NeighborNet* algorithm represents linguistic relationships as a network, allowing for the visualization of conflicting or non-hierarchical signals (Bryant and Moulton, 2004; List et al., 2014). This is particularly important for capturing continuous variation in acoustic features, which may not conform to a strictly tree-like pattern and could reflect contact, convergence, or gradual divergence processes. Figures 5.3A & B illustrate the neighbor-nets derived from population-level vowel acoustics (VoxCommunis data; Ahn and

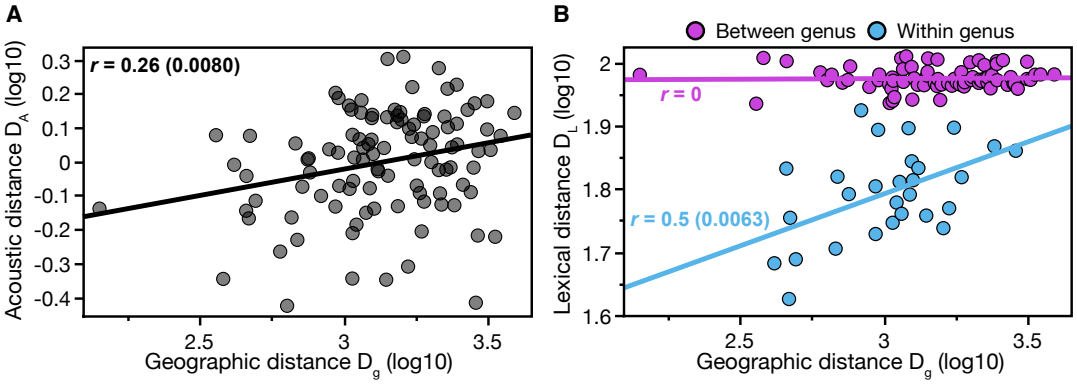


FIGURE 5.2: **Isolation by distance in acoustic and lexical data**. **A**: Acoustic distance (D_A) between language communities vs. geographic distance (D_g). **B**: Correlation between geographic distance (D_g) and lexical distances (D_L). The data is grouped for between-genus and within-genus comparisons for the different genera (Slavic, Romance, Germanic, Isolate) of the Indo-European languages studied here.

Chodroff, 2022) and lexical distances (ASJP database via Levenshtein distance; Wichmann et al., 2022), respectively.

Together, the two neighbor-nets reflect distinct structural signals in the data: lexical distances primarily capture genealogical clustering, while acoustic distances reveal both genealogical grouping and influence from geographic proximity and contact. The acoustic neighbor-net broadly clusters languages by genus, with Romance, Slavic, and Germanic languages generally grouping together. Some deviations from known genealogical relationships reflect known cases of language contact (compare Fig. 5.1D). For instance, Basque, although a language isolate, clusters acoustically close with Spanish due to regional convergence through bilingualism and contact. Similarly, Sorbian and Polish, both Slavic languages, show acoustic proximity to Germanic varieties due to intense contact and bilingualism.

In contrast, the lexical neighbor-net (Fig. 5.3B) more rigidly reflects genealogical relationships. Languages cluster cleanly within their respective Indo-European genera, and Basque is a clear outlier, showing no lexical similarity to Indo-European languages. Notably, Romance and Slavic subgroups form tightly supported clusters (e.g., Polish–Sorbian, Italian–Romanian), while Germanic languages cluster separately.

5.3.3 Does acoustic variation retain deep historic signals?

The neighbor-net comparisons revealed that while lexical data closely mirror genealogical lineages, acoustic data reflect a more complex interplay of inheritance and interaction. This raises a broader question: beyond contact and recent history, can acoustic variation preserve deep historical signals of language divergence? Given that isolation by distance (IBD) effects for lexical distances are only observed in language genera of the Indo-European family, but not across more divergent language genera, this disparity suggests that while geographic proximity contributes to lexical similarity among closely related languages, it becomes less predictive at broader phylogenetic scales, such as between language families (Starostin, 2000; Gray and Atkinson, 2003; Gray, 2005; Heggarty et al., 2023). The turnover and replacement of lexical items over time erase historical continuity (Renfrew et al., 2000), limiting the depth at which lexical data can reveal linguistic and population history (Blythe, 2015; Greenhill et al., 2017). These patterns are consistent with models of language

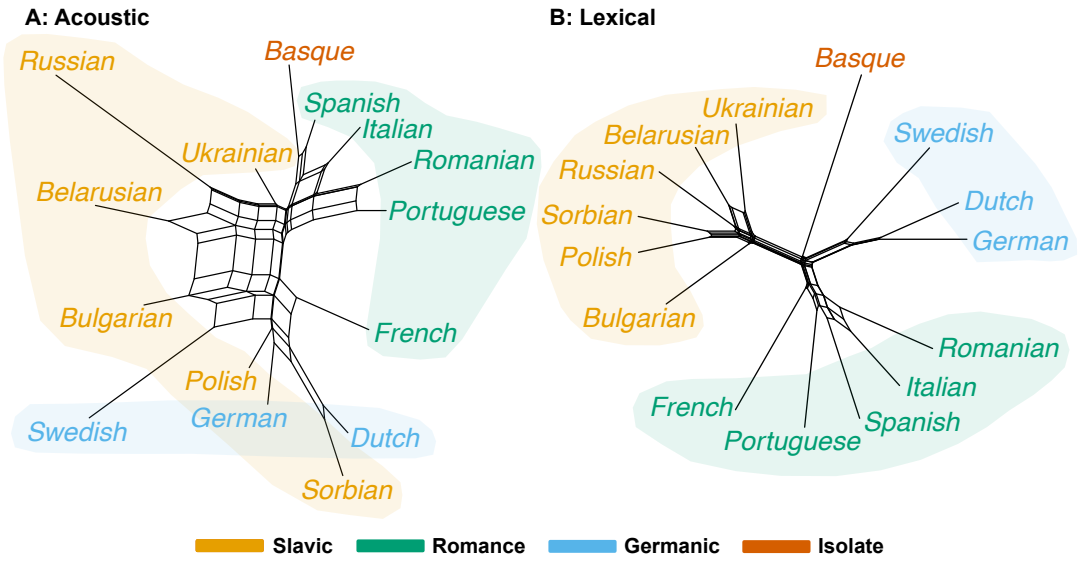


FIGURE 5.3: **Comparison of acoustic and lexical neighbor-nets**. **A**: Neighbor-net based on neighbor-joining of acoustic distances between speech communities (see Fig. 5.1C). Languages are colored according to their genus (Slavic, Romance, Germanic, Isolate) in the Indo-European language family. Note that Basque is a language isolate. **B**: Neighbor-net based on neighbor-joining with lexical distances (Levenshtein distance) from the ASJP data base (Wichmann et al., 2022).

change that emphasize punctuated bursts of evolution, suggesting that lexical systems may undergo rapid shifts during population splits, migrations, or other major demographic events (Atkinson et al., 2008).

In contrast, vowel space acoustic variation shows a consistent isolation by distance (IBD) pattern across the full language sample, indicating a gradual, drift-like trajectory of change over time and space – consistent with neutral evolution models (Moran et al., 2024). Just as IBD has been used in population genetics to reconstruct population structure and infer human migratory history (Cavalli-Sforza et al., 1994; Relethford, 2004; Novembre et al., 2008), we propose that IBD in vowel acoustics can offer a powerful tool for investigating deep-time questions in historical linguistics. The IBD pattern persists even where contact-induced convergence is evident (Figs. 5.1D and 5.3A). Moreover, clustering based on vowel acoustics broadly reflects known genealogical relationships (Fig. 5.3), suggesting that vowel systems can retain historical signals. Overall, these findings offer a complementary perspective to punctuated models of lexical change (Atkinson et al., 2008), supporting the view that different components of language evolve in distinct modes and tempos (Parravicini and Pievani, 2018).

Similar to how acoustic signals have been used to recover phylogenetic signals in other vocal learners such as birds (Mason et al., 2017; Arato and Fitch, 2021; Rivera et al., 2023), our analyses suggest that human acoustic variation also retains historical information that transcends recent cultural change. As such, acoustic data may prove especially valuable in reconstructing broad-scale between-language and between-language family relationships and in testing long-range linguistic history hypotheses. In particular, acoustic vowel system divergence – analyzed at the level of individual speakers and speech communities – could be leveraged to explore large-scale hypotheses such as the Out-of-Africa dispersal of modern humans and subsequent global patterns of linguistic diversification (Atkinson, 2011).

Overall, our results highlight the distinct evolutionary dynamics of lexical vs

phonetic systems. While lexical structure closely tracks relatively recent genealogical relationships and language contact, acoustic features reveal a more gradual process of divergence, shaped by neutral-like dynamics (Moran et al., 2024), geographic diffusion and contact situations. This divergence in evolutionary dynamics reinforces the importance of multi-component analyses in language evolution research (Parravicini and Pievani, 2018), which underscores the need to integrate phonetic, lexical, and genetic data in comparative studies of language evolution (Barbieri et al., 2022). Future research should expand these approaches to additional language families and regions to assess the universality of these dynamics and refine our understanding of deep-time linguistic change.

5.4 Materials and Methods

Data

We obtained acoustic-phonetic data via the pipeline available from the in the Vox-Communis Corpus (Ahn and Chodroff, 2022), an extension of Mozilla Common Voice (Ardila et al., 2020). We selected N=14 Indo-European languages and Basque as a language isolate (Basque, Belarusian, Bulgarian, Dutch, French, German, Italian, Polish, Portuguese, Romanian, Russian, Sorbian, Spanish, Swedish, Ukrainian) based on the availability of a sufficient number of speakers. We included Basque in our analysis despite it being a non-Indo-European language because it provides a valuable outlier for understanding patterns of linguistic variation (see main text).

The selected languages include the three cardinal vowels (i, a, u) necessary to construct vowel triangles using the formants F1-F2. Only speakers who produced the complete set of these vowels were included, resulting in N=55,116 vowel utterances from N=18,372 individuals. Further details on the acoustic data samples, including sample sizes per population, can be found in Table 5.1. Sample sizes varied substantially across languages, primarily reflecting differences in data availability within the Common Voice corpus (Ardila et al., 2020). To ensure that this variation did not bias our results, we tested whether within-group acoustic variation was correlated with sample size per language. Specifically, we regressed the log10 of within-group acoustic variance against the log10 of sample size. No positive trend was observed ($R^2 = 0.03$, $p = 0.5$), suggesting that our sampling captures inherent population-level variation rather than artifacts of uneven representation. This supports the reliability of our dataset for cross-linguistic comparisons of vowel system structure. The lexical distance data is from the ASJP database and was used to assess the lexical distance between the sampled languages (Wichmann et al., 2022).

Statistical analysis

All statistical analyses were conducted in *R* (R Core Team, 2024). Formant data were converted to the Bark scale to align with human auditory perception (Traunmüller, 1990) (see Fig. 5.1B). We did not apply speaker normalization techniques which typically remove most interindividual variation (Adank et al., 2004), as recent work suggests that such procedures may obscure meaningful inter-speaker and inter-population variation (Josserand et al., 2024; Moran et al., 2024). Retaining this inherent variation is essential for capturing the full range of acoustic diversity within and between populations, which in turn is necessary for investigating the dynamics of evolutionary processes.

Each speaker's vowel system was represented as a vector of length 6 (3 vowels × 2 formants: F1, F2). Principal component analysis (PCA) was applied to the resulting 18,372 × 6 matrix (individuals × acoustic dimensions), yielding a low-dimensional representation of the vowel space (Fig. 5.1C). To compare within- and between-population acoustic variation, we computed the variance over the first six PCs for both individual-level data (within-population: $\text{Var}_{\text{w.lang}}$) and population means (between-population: $\text{Var}_{\text{b.lang}}$), and report their ratio ($\text{Var}_{\text{w.lang}}/\text{Var}_{\text{b.lang}}$) in Table 5.1.

Acoustic distances between populations were calculated as Euclidean distances between population means in the PCA space (PCs 1–6). Lexical distances were computed analogously, using average Levenshtein distances from the ASJP database (Wichmann et al., 2010). Geographic distances between speech communities were derived from coordinates provided by the GeLaTo database (Barbieri et al., 2022), and represent geodesic distances computed with the R library *geosphere* (Hijmans, 2024). All distance data were log10-transformed and vectorized to assess relationships among acoustic, lexical, and geographic distances via linear regression (see Fig. 5.2). Log10 transformation is necessary here to account for the skewed distribution of pairwise distances and to linearize relationships, which is standard in studies investigating isolation by distance (Slatkin, 1993).

To assess structural similarity between acoustic and lexical data, we constructed neighbor-nets from the respective acoustic and lexical distance data using the R package *Phangorn* (Schliep, 2011) (Fig. 5.3). Unlike traditional tree-based methods, the *Neighbor-Net* algorithm represents linguistic relationships as a network, allowing for the visualization of conflicting or non-hierarchical signals. We apply *Neighbor-Net* to both acoustic and lexical data to enable a consistent comparison across domains, while accommodating the possibility that acoustic diversification may not follow a strictly tree-like pattern (Bryant and Moulton, 2004; List et al., 2014).

Conclusions

At the outset of this dissertation, a central question was posed: how does inter-speaker variation in the morphology of the speech organs, in articulatory realizations, and in the resulting acoustic output contribute to the diversification of languages, and ultimately to the vast linguistic diversity observed across the globe? Throughout the various chapters, different approaches were undertaken to explore this question. As is often the case in science, the answers uncovered here address parts of the original question, but more importantly, they open up new avenues for inquiry, generating fresh hypotheses and prospects for future research. Thus, the study of *morpho-phonetics* and the methods applied in this dissertation are far from exhausted; indeed, this work may only mark the beginning.

This dissertation has demonstrated in detail how the extensive interindividual variation in the morphology of the speech organs shapes and conditions the acoustic variation of vowels within speaker groups, leaving systematic acoustic signatures. Furthermore, it has been shown that vocal-tract-conditioned acoustic variation contributes to the diversification of vowel acoustics across languages (Chapter 2).

Additionally, the *morpho-phonetic* investigation of coarticulation in vowel-consonant-vowel (VCV) sequences reveals that while consonants and vowels shift their locations in multivariate articulatory and acoustic spaces under coarticulatory influences, they maintain their relative contrasts – emphasizing that speech production prioritizes relational rather than absolute articulatory and acoustic targets. Thus, this work provides additional insights into the dynamics of coarticulation and the emergence of articulatory and acoustic variation in natural speech (Chapter 3).

Building on these observations, Chapters 4 and 5 examined how spoken languages and birdsong diversify acoustically, revealing striking parallels between the acoustic diversification of language and birdsong, and equally striking differences in how the acoustics and lexicons of languages diversify. Taken together, these analyses suggest that neutral-like evolutionary processes play a major role in the acoustic diversification across species of vocal learners. Moreover, a consistent pattern of isolation by distance across speaker and songbird populations indicates that acoustic data may hold promise for reconstructing the origins and migration histories of languages and populations.

The following sections will expand on the aspects outlined above, situate them within the broader scientific context, and draw inferences for future research directions in this field.

6.1 The dawn of geometric morphometrics in speech science

To the best of my knowledge, this dissertation represents the first application of the full methodological toolbox of geometric morphometrics (GM) to the speech sciences. Previous work was either preliminary in nature (Zollikofer et al., 2011; Gully, 2021), or did not leverage the full anatomical scope of the vocal tract (Popat et al.,

2013), nor fully integrate the Procrustes paradigm (Mooshammer and Geng, 2008; Geng and Mooshammer, 2009). However, the research possibilities in applying GM to the speech sciences are vast.

6.1.1 A new perspective on the vowel polygon

This dissertation is the first work to provide a comprehensive representation of the articulatory vowel space, i.e., the vowel polygon, using geometric morphometrics, based on real-time MRI articulatory data recorded during speech. Previous preliminary articulatory mappings of vowel spaces were achieved using EMA (Electromagnetic Articulography) data or with modeling approaches (Weirich and Simpson, 2014; Xu et al., 2021). The advantage of representing the articulatory vowel space using landmark-based geometric morphometrics lies in its ability to capture both a low-dimensional shape space (via PCA) and the corresponding physical morphologies simultaneously (see Fig. 6.1). Additionally, GM allows for a holistic multivariate representation of the vowel space, enabling the analysis of all articulators involved in speech production at once (Xu et al., 2021). Given the long-standing discussions surrounding the mapping between articulation and acoustics (Blumstein and Stevens, 1979; Browman and Goldstein, 1992; Galantucci et al., 2006; Neiberg et al., 2008; Noiray et al., 2014; Perkell and Klatt, 2014), the establishment of an articulatory vowel space grounded on standard multivariate statistics can offer valuable insights into how vocal tract morphology shapes speakers' articulatory strategies (Fuchs et al., 2008), and, consequently, their acoustic output. In Chapters 2 and 3, it was demonstrated that the articulatory vowel space closely parallels the acoustic vowel space (Moran et al., 2024) and correlates with acoustic spaces capturing formant representations of vowels as well as the spectral envelope of speech sounds (see Chapter 3). As neuroimaging and speech perception research has identified articulatory and acoustic spatial representations in the brain (Chang et al., 2010; Scharinger et al., 2011; Bouchard et al., 2013a; Mesgarani et al., 2014; Bouchard et al., 2016; Cheung et al., 2016; Chartier et al., 2018; Preisig et al., 2022; Oganian et al., 2022), this work suggests that future studies should pursue integrated analyses of articulation, acoustics, and perception to achieve a more comprehensive understanding of the entire speech production-perception loop (Cheung et al., 2016).

The establishment of a detailed articulatory representation of the vowel space further offers new opportunities to explore the evolution of vowel systems, that is, the composition of speech sounds found in the world's languages today. The distribution of speech sounds in today's languages, including vowel inventories, is by no means random, as proposed by various hypotheses and theories on the emergence of vowel systems (Lindblom, 1990a; Lindblom, 1990b; Lindblom, 1990c; Schwartz et al., 1997; De Boer, 2000; De Boer, 2001; Vaux and Samuels, 2015; Zhang and Gong, 2022). One prominent line of thought explains the evolution of vowel systems through principles of self-organization. According to this view, vowel systems emerge under functional constraints on articulation and acoustics, and through learning processes that optimize efficiency. Computer simulations have demonstrated that such conditions can lead to the self-organized formation of vowel systems approaching "optimal" configurations (De Boer, 2000; De Boer, 2001).

Other approaches emphasize the role of perceptual constraints in shaping vowel systems. In these accounts, vowel systems evolve to maximize vowel *dispersion* – ensuring that vowels are sufficiently separated in perceptual space to maintain distinct contrasts – and *focalization* – favoring vowels with closely spaced formants, which facilitates easier perception (Liljencrants and Lindblom, 1972; Schwartz et al., 1997;

Lindblom, 1990a). This also ties to other work, that highlights constraints related to the *learnability* of vowel systems: systems that are easier to acquire – for example, those with more widely dispersed sounds in perceptual space – are predicted to become more prevalent across languages over time compared to harder-to-learn systems (Vaux and Samuels, 2015).

Finally, several theories also recognize the role of biological processes, particularly the anatomy of the vocal tract, in shaping vowel systems (Lindblom, 1990a; Lindblom, 1990c). Although self-organization models incorporate some aspects of articulatory constraints (De Boer, 2000; De Boer, 2001), direct empirical evidence for the specific influence of vocal tract morphology on the evolution of vowel systems remains limited or underexplored (Lindblom, 1990c) – especially in light of the vast vocal tract variability between speakers identified in this work. Therefore, a robust and informative representation of the articulatory vowel space, as developed in this work, may contribute significantly to addressing this gap in our understanding.

6.1.2 The morphospace of vowel production

One approach to studying the longstanding and venerable notion of the *possible speech sound* – that is, the *total sound-producing potential* of humans from an articulatory perspective (Catford, 1977; Lindblom, 1990c), is through the investigation of morphospaces (Budd, 2021). Building on the foundational work of Raup's famous morphospace of coiled shells (Raup, 1966), the analysis of articulatory data with geometric morphometrics (GM) allows for the exploration of the morphospace of vowel production (Fig. 6.1), and could thus offer insights into the range of possible speech sounds shaped by vocal tract morphology. David Raup originally mapped the shapes of shelled invertebrates within a low-dimensional configuration space defined by their mathematical parameters, exploring whether their evolutionary forms are limitless, or constrained to a specific set of shapes (Raup, 1966; Budd, 2021). A morphospace can thus be understood as a type of configuration space, where entities – here, vowels defined by their articulatory configurations – are positioned according to selected morphological features. In this study, these features are captured as the statistically independent principal components of shape (or "shape components"), which in turn relate to the movement of the articulators to produce vowel sounds. Thus, morphospaces provide a framework for systematically mapping the shapes of individual objects (such as vowels) based on their deviation from an average or consensus shape – which is precisely what GM and PCA of shape variables quantify (Bookstein, 2018). This approach allows for rigorous, quantitative comparisons of object positions within the morphospace. Two key metrics commonly analyzed are *occupancy*, which refers to the portion of the morphospace occupied by specific groups, and *disparity*, which measures the degree of morphological differentiation between groups or taxa. Together, these metrics enable inferences about evolutionary constraints in developing systems (Budd, 2021). In the final part of this dissertation – to illustrate this principle – a morphospace of vowel production was constructed using articulatory data from Chapter 2, visualized in a low-dimensional (3D) shape space (Fig. 6.1).

Here, some findings on the exploration of the "morphospace of vowel production" are presented. Figure 6.1 shows that both articulatory and acoustic multivariate spaces may be visualized as low-dimensional (here: 3D) feature spaces (Fig. 6.1A–B), and their close resemblance indicates that these spaces correspond to each other, as was shown in Chapters 2 and 3. Shape changes along the three dimensions of the articulatory shape space correspond to articulatory changes in the vocal

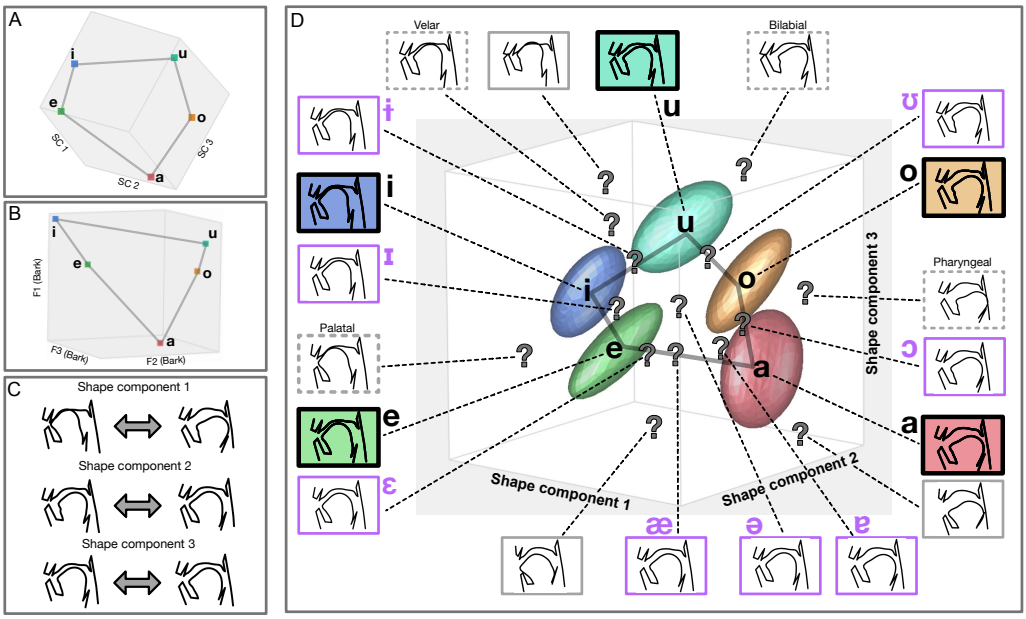


FIGURE 6.1: **The morphospace of vowel production. A**: The articulatory vowel polygon represented in 3D shape space by the first three shape components (SC1–SC3), with the 3D cube rotated to best match the conventional IPA vowel chart design (International Phonetic Association, 1999). **B**: The acoustic vowel polygon based on the first three formants (F1–F3) plotted in the Bark scale (Traunmüller, 1990). **C**: Articulatory shape changes associated with SC1–SC3, shown at the extreme ends of variation. **D**: Morphospace of speaker-normalized vowel articulation, where vowel symbols indicate the mean articulatory position for each cardinal vowel with 50% normal contour ellipsoids representing interindividual variation for each vowel (N=32 speakers). Colored boxes show the actual mean articulatory shapes of the cardinal vowels. Purple boxes indicate inferred vowel shapes (marked with "**?**)" based on their expected positions within the IPA vowel polygon, projected via shape interpolation within the cardinal vowel morphospace. Grey boxes show interpolated articulatory shapes located outside the typical IPA vowel area; notably, some shapes within dashed boxes resemble articulations of certain consonant groups, such as palatal, velar, bilabial and pharyngeal configurations. The 3D visualizations in this figure were created using JMP Pro 18.0.2

tract, such as the antero-posterior position of the tongue, height of the tongue as well as lip opening and closing (Fig. 6.1A & C). The articulatory configuration in 3D space resembles the classic IPA vowel chart (Fig. 6.1A), which was originally designed to bear an approximate relation to the tongue position during vowel articulation (International Phonetic Association, 1999). The data presented here reveal that the original assumptions were not far off: the relationships proposed by the IPA chart closely align with those observed in articulatory data. However, the approach taken here captures these relationships quantitatively and more comprehensively, going beyond tongue position to include other articulatory structures such as the soft palate and the lips.

The exploration of the morphospace of vowel production suggests that the organization of the vowel space is substantially influenced by vocal tract morphology and its constraints. First, interindividual variation in vowel production appears to be structured largely orthogonal to between-vowel variation (Fig. 6.1D), meaning that more vowels can "fit" into the morphospace because the space between vowels is not densely occupied by inter-speaker variation. This non-correlation may

arise from two, not mutually exclusive, mechanisms: (1) vowel systems may have evolved along dimensions that are orthogonal to the principal axes of interindividual vocal tract variation, thus maximizing phonemic discriminability; and (2) individuals may acquire or adjust their vowel productions to minimize the influence of anatomical variation on vowel identity. Second, since the distances traveled between vowel categories in morphospace are proportional to articulatory distances (i.e., the extent of articulatory change required between vowels), the morphospace of vowel productions suggests a limitation on how many distinct vowels can fit between two given vowels – not only for perceptual distinctiveness (cf., Liljencrants and Lindblom, 1972; Lindblom, 1990a; Schwartz et al., 1997; Vaux and Samuels, 2015), but also for articulatory feasibility and simplicity (Lindblom, 1990c). This observation indicates that articulatory distances in morphospace may be proportional to perceptual distances, implying a feedback mechanism between articulation and perception to ensure *sufficient contrast* between speech sounds (Lindblom, 1990c; Browman and Goldstein, 1992; Galantucci et al., 2006).

Moreover, the morphospace of vowel production highlights obvious articulatory vocal tract constraints: certain areas of the morphospace remain unoccupied, and interpolated shapes in these regions (visualized through GM methods; Bookstein, 2018) often correspond to articulatorily implausible configurations (Fig. 6.1). These configurations either produce constrictions characteristic of consonantal articulation (e.g., a tightly closed pharyngeal constriction or complete tongue-palate contact), or result in spatially constrained, non-viable shapes – such as an extremely retracted tongue root – rendering them articulatorily implausible (see gray boxes in Fig. 6.1). Readers are encouraged to attempt the articulatory gesture depicted at the bottom right of Fig. 6.1D (below the vowel rendering of [a]), to experience this limitation firsthand. Interestingly, some of these "impossible" vowel shapes resemble known consonantal articulations (see gray dashed boxes in Fig. 6.1), suggesting that regions unoccupied by vowels in articulatory space may correspond to areas populated by consonants – a hypothesis warranting further investigation. In contrast, interpolated vowel shapes that lie within or along the boundaries of the typical vowel morphospace resemble plausible vowel configurations (see purple boxes in Fig. 6.1).

Overall, these findings suggest that the structure of the articulatory vowel space – and, consequently, the acoustic vowel space – is shaped not only by perceptual constraints but also significantly by morphological ones. While theories of self-organization in vowel systems have illuminated how structure may emerge from transmission and perceptual optimization (De Boer, 2000; De Boer, 2001), the present findings suggest that such processes are significantly constrained by the anatomical and biomechanical structure of the vocal tract. The vowel morphospace is not evenly accessible, and certain regions are articulatorily implausible, challenging the assumption of a uniformly navigable space often used in self-organization models. Future models should incorporate such morphological constraints to more accurately reflect the realities of speech production. This work therefore proposes that articulatory morphospaces should be explored more comprehensively, extending beyond vowels to include consonants, to better understand the full extent of possible speech sounds. Such an approach could yield new insights into the emergence of vowel systems, tying back to Raup's original ideas about morphospace (Raup, 1966). Just as coiled shell forms are constrained by *occupancy* (organisms can only occupy viable regions) and *disparity* (variation between forms/shapes), the vowel morphospace appears similarly shaped by the need for vowels to be both articulatorily possible (occupancy) and sufficiently distinct for their perception (disparity).

6.1.3 The impact of coarticulation on speech and language

In Chapter 3, it was demonstrated that coarticulation significantly affects both the articulation and acoustics of vowels and consonants. The main findings from that chapter are that both vowels and consonants rely on contrasts between speech sounds for their production, acoustic maintenance and perception i.e., the relative positioning of sounds to each other in articulatory and acoustic space plays a crucial role. This resonates with theories of vowel dispersion, which propose that vowel systems are organized to maximize perceptual distinctiveness between categories (Liljencrants and Lindblom, 1972; Lindblom, 1990a; Schwartz et al., 1997; Vaux and Samuels, 2015). In this context, coarticulation can be seen as a shaping force that both constrains and reinforces dispersion patterns by influencing how reliably contrastive distances can be maintained across varying phonetic environments. Additionally, the investigations into coarticulation reveal that vowels tend to be more stable (or constrained) compared to consonants. This stability may stem from the fundamental difference in how consonants and vowels are produced. Consonants require a constriction produced within the vocal tract (for example, producing the sound [k] clearly demonstrates this), whereas vowels involve the shaping of the vocal tract by the articulators (like the tongue) to produce specific vocal tract contours (as discussed in Chapter 1.5; Fant, 1960). Therefore, consonants are less prone to acoustic alterations that have their origin in regions of the vocal tract that do not affect their constriction (Chapter 3).

These findings connect to the morphospace paradigm outlined in Section 6.1.2. Since coarticulation leads to context-dependent shifts in vowel production, and given the perceptual and articulatory constraints discussed in Section 6.1.2, vowels may only shift within certain bounds due to coarticulation. These shifts must ensure that vowels still occupy a feasible position in the morphospace, and more importantly, remain perceptually distinct as the intended target sound (Lindblom, 1990c). If vowels shift too far from their target, this poses the risk that they are being perceived as a different speech sound altogether. However, the research in Chapter 3 also demonstrated that speech sounds compensate for coarticulatory effects (Tilsen, 2013), mainly through their relative positioning in both articulatory and acoustic spaces. The relative positioning of sounds is crucial for both production and perception, but this compensation effect likely operates within certain boundaries, as shown in Fig. 6.1D. Beyond these boundaries, a coarticulation-induced shift may lead to misinterpretation or reinterpretation of the target sound. Such instances could serve as catalysts for sound change (Beddor, 2009), as coarticulation introduces synchronic variation into the speech stream, potentially acting as a precursor to sound change (Ohala, 1989).

Furthermore, coarticulation may facilitate (or has facilitated) to "fill in the gaps" within the morphospace of possible speech sounds, suggesting a connection between coarticulation and phonology, i.e., the sound systems of languages (Ohala, 1993). Since coarticulation causes shifts in both vowels and consonants, vowel-consonant coarticulation may have governed the exploration of the vowel space, potentially leading to the development of new vowel categories within the morphospace based on the specific consonantal inventory of a language. In summary, these observations suggest that the consonant set in a language not only constrains the set of vowels that may be produced but may have also enabled the exploration of new vowels within the morphospace – within perceptually salient boundaries. This raises an interesting question: how does the consonant system of a language influence its vowel system, and vice versa? Do current sound systems of languages

retain a historical signal reflecting the constraints vowels and consonants impose on each other? This ties back to older questions which asked whether there are universal constraints in sound sequences in the worlds languages (Kawasaki and Ohala, 1980; Janson, 1986). Given the stability of vowel systems, which can only shift so much due to coarticulation while maintaining contrast (see Chapter 3), it is possible that vowels influence consonantal configurations. If certain coarticulatory patterns compromise contrast, some sounds may become unfeasible. In contrast, the consonant system may constrain vowel qualities, as certain vowels may be less likely in specific consonant contexts. This suggests the potential of a systematic relationship between vowel and consonant systems, where the presence of certain consonants might predict vowel patterns, and vice versa (e.g., phonotactic constraints). There is some research that supports these claims, particularly in certain sound sequences (Kawasaki and Ohala, 1980; Janson, 1986; Rousset, 2003; Doucette et al., 2024). Given the ongoing debate and lack of consensus on these questions (Maddieson and Precoda, 1992), future research – especially employing geometric morphometrics (GM) and morphospace analysis – could provide further insights into these interactions.

Overall, these findings thus raise the possibility that coarticulation plays an important role not only in speech production but also in the evolution of phonological systems (Ohala, 1993; MacNeilage and Davis, 2000). This relationship between coarticulation and vowel space organization may be investigated using an *evo-devo* framework (Müller, 2007), particularly through developmental studies in children (Vihman, 2014). Since coarticulatory patterns are prominent in early speech and gradually stabilize as the vocal tract and neural control matures (Namasivayam et al., 2020), the emergence of vowel categories in ontogeny may reflect how articulatory biases – including constraints from coarticulation – shape phonological inventories.

6.1.4 Reconstructing speech capacity using GM

Addressing the complex interactions outlined in the previous sections also ties into broader questions about the origins and evolution of language itself. The origin of language and the emergence of the capacity for speech are widely debated topics (Hauser et al., 2002), particularly regarding the question of when these abilities first arose (Hauser et al., 2014). Since spoken language does not fossilize – unlike writing, which dates back to roughly 5,000 years before present (Schmandt-Besserat, 2014) – its reconstruction is inherently challenging. Additionally, the comparative method in historical linguistics – based on systematic comparison of language cognates to reconstruct their genealogical relationships and vocabulary – has an generally agreed-upon time depth of around 8,000 to 10,000 years ago (Nichols, 1992; Gray, 2005).

Over the past 40 years, discussions have centered on whether Neanderthals possessed a vocal tract comparable to that of modern humans, whether their speech capabilities were similar to ours, and how large their vowel spaces might have been (Lieberman et al., 1972; Lieberman and Crelin, 1972; Boë et al., 2002; Dediu and Levinson, 2013; Janssen et al., 2019). Genetic evidence has particularly focused on the gene *FOXP2*, a transcription factor involved in the neural circuits underlying language. Its accelerated evolution in modern humans around 260,000 years ago – resulting in a variant that differs from that of the Neanderthals – initially fueled the hypothesis of enhanced language capabilities in modern humans compared to other hominins (Konopka et al., 2009; Maricic et al., 2013). However, more recent evidence

suggests that Neanderthals may have had auditory and speech capacities, and possibly even language abilities, similar to those of modern humans (Johansson, 2015; Conde-Valverde et al., 2021).

Consequently, estimates for the origin of speech and language are widely diverging. The discontinuity theory posits a rapid emergence of language around 50,000 to 100,000 years ago (Lieberman, 2007; Bolhuis et al., 2014; Berwick and Chomsky, 2016). In contrast, more recent studies suggest that the roots of language may trace back to the common ancestor of modern humans, Neanderthals, and Denisovans, approximately 500,000 years ago (Dediu and Levinson, 2013; Dediu and Levinson, 2018). Meanwhile, the continuity theory argues for a gradual evolution of language over millions of years (Pinker and Bloom, 1990; Ulbaek, 1998; Lieberman, 2015; Barham and Everett, 2021). However, the concrete origin and timing of speech and language remain largely enigmatic (Bolhuis et al., 2014), as existing estimates are uncertain and reconstruction methods rely on diverse, and often conflicting, lines of evidence without a clear consensus. A major complicating factor is that speech and language depend not only on the anatomy of the vocal tract but also on sufficient neural capacities (Fitch et al., 2016). Thus, even if the vocal tracts of fossils appear anatomically compatible with speech production, the cognitive prerequisites must also be taken into account (Dubreuil and Henshilwood, 2013; Ponce de León et al., 2021).

One promising avenue for reconstructing the vocal capabilities of ancient humans (such as Neanderthals or *Homo erectus*) lies in the reconstruction of their vocal tracts. Fossilized skeletal structures such as jaws and hard palates can, in principle, be used to infer associated soft tissue structures, and even the vocal capacities through computational modeling and simulation (Dediu et al., 2021). Geometric morphometrics (GM) and related advanced methods offer tools to reconstruct soft tissue based on skeletal data (Lautenschlager, 2016; Landi and O'Higgins, 2019; Dediu et al., 2021), potentially providing valuable new insights into this challenging area. Furthermore, sophisticated *morpho-phonetic* analyses based on GM could also offer critical perspectives on long-standing assumptions regarding the uniqueness of the modern human vocal tract compared to that of monkeys and apes – and how this vocal tract anatomy relates to its function in producing complex speech sounds (Lieberman et al., 1972; Fitch, 2000; Fitch and Reby, 2001; Fitch et al., 2016; Janssen et al., 2019).

While the analysis of vocal tract imaging data presented here offers some insights into this direction, it remains limited by its focus on two-dimensional representations, specifically within the midsagittal plane of MRI images. A more holistic approach to studying articulation and vocal tract morphology using GM – particularly for reconstructing ancient vocal tracts – should incorporate three-dimensional data. Fortunately, several methods for extracting and visualizing 3D vocal tract structures already exist, and a growing number of 3D data resources are available (Narayanan et al., 2014; Fu et al., 2017; Lim et al., 2019; Birkholz et al., 2020; Isaieva et al., 2023; Erattakulangara et al., 2025). Thus, future research should aim to extend *morpho-phonetic* methods into three dimensions, enabling a more holistic understanding of speech production mechanisms and advancing the potential for reconstructing the vocal tracts of ancient humans.

In summary, it is proposed here that future research into the origins of speech capacity should leverage advanced methods of geometric morphometrics alongside concepts discussed throughout this work, such as the morphospace of vowel production (see Section 6.1.2). Such an integrated approach could offer valuable insights into how ancient humans may have produced speech. However, and crucially, since

there is still no unified understanding regarding how vocal tract dimensions and articulation exactly map onto the acoustics of speech (although this dissertation has provided some insights into this gap), it is essential that future research continues to address this issue. Reconstructions of vocal capacities must be grounded in stable models and well-supported assumptions about the relationship between articulation, acoustics, and perception.

6.2 Studying variation is key in language evolution research

Having explored how geometric morphometrics and morphospaces could illuminate the evolution of speech production, it was also shown in Chapters 2, 4 and 5 how variation in contemporary languages – arising in part from both anatomical and articulatory differences – contributes to ongoing language diversification. This perspective emphasizes the importance of inter- and intra-speaker variation as a substrate for linguistic evolution. In this work, various processes that introduce variation into the speech signal have been discussed, specifically those related to the articulation of speech sounds, which subsequently affect the acoustic domain. The key sources of variation identified in this study are interindividual vocal tract differences and articulatory strategies, as well as coarticulation between vowels and consonants.

This variation can be interpreted through different perspectives. It may be viewed as noise within the speech signal, blurring linguistic distinctions between individual speakers or speaker groups, as is the common view in various language research disciplines that normalize these differences across speakers (Labov, 1966; Lobanov, 1971; Adank et al., 2004; Giuliani et al., 2006; Burridge, 2023). Alternatively, it can be seen as valuable information, i.e., meaningful variation in the acoustic signals between speakers that may drive diversification processes (Moran et al., 2024). Evolutionary theory posits that for new traits (such as acoustic traits or speech sounds) to emerge, there must be heritable variation (here: heritability through cultural transmission) upon which adaptive or non-adaptive evolutionary processes can act (Lewontin, 1970). In other words, interindividual variation is crucial for understanding evolutionary processes (Lande, 1980; Lande and Arnold, 1983; Lande, 1991; Ackermann and Cheverud, 2004) – a principle that has been suggested to similarly affect the diversification of language (Ohala, 1989). However, this idea has received limited attention, as historical linguistics – which is particularly rooted in the comparative method – has dominated the field of language diversification (Atkinson and Gray, 2005; Campbell and Poser, 2008; Campbell, 2013), providing significant insights into language diversification and the (widely debated) origin of language (Gray and Jordan, 2000; Gray and Atkinson, 2003; Atkinson et al., 2008; Atkinson, 2011; Greenhill et al., 2017; Heggarty et al., 2023). The comparative method analyzes the discrete and hierarchical properties of languages, such as sound inventories or lexical cognates (words inherited from a common historical origin), to infer the diversification and genealogical relationships between languages (Atkinson and Gray, 2005; Campbell, 2013; Greenhill et al., 2017; Barbieri et al., 2022). However, as previously mentioned, the comparative method is limited by a temporal constraint (approximately 8,000–10,000 years before the present), due to the compounding effects of lexical turnover, which can dilute the historical signal, particularly in lexical comparisons (Gray, 2005).

Furthermore, an important distinction must be made. While comparative methods, including phylogenetic studies, reveal the historical (evolutionary) relationships between languages or species, these methods offer limited potential for uncovering the evolutionary processes at the level of individuals and populations that led to such divergence (Mooers and Heard, 1997; Uyeda et al., 2018). As originally stated by Charles Darwin (Darwin, 1859), to disentangle the concrete evolutionary processes driving diversification, one must examine variation across individuals and groups. While the comparative method will continue to provide valuable insights into language diversification – as has recently been demonstrated with the integration of genetic data in combination with phylogenetic reconstruction of languages (Pakendorf, 2014; Barbieri et al., 2022) – I argue for the further diversification of methods (such as evolutionary population biology) and the integration of additional lines of evidence in the study of language evolution, particularly focusing on acoustic variation. Chapters 2, 4 and 5 have provided detailed explanations on how language diversification can be studied through acoustic data.

6.2.1 The *morpho-phonetic model of acoustic trait evolution*

In summary, the approach described in Chapters 2, 4 and 5 is here unified under the *morpho-phonetic model of acoustic trait evolution*. Note that acoustic traits are generalized here, but may refer to the acoustic measurements of vowels, the full vowel polygon (as shown in Chapters 2, 4, 5), or any other form of acoustic measurement. This model builds on the classical population biology framework, which asserts that the variation in a population's phenotype (V_P) is proportional to the sum of genetic variance (V_G) and environmental variance (V_E) (Falconer, 1996):

$$V_P \sim V_G + V_E \tag{6.1}$$

In other words, the observed phenotypic variation in a group can be explained by the combination of its genetic and environmental variance. For the purposes of modeling acoustic trait evolution, we adapt this model by assuming that *phonotypic* variation (V_{Phon}; variation in acoustic traits) is proportional to the additive effects of variation in the phenotype (V_P; variation in the vocal tract) and environmental variance (V_E):

$$V_{Phon} \sim V_P + V_E \tag{6.2}$$

Thus, we repurpose the vocal tract (VT) as the "phenotype" and the acoustic traits under study as the "phonotype" in analogy to the original model. This adapted model can therefore be rewritten as:

$$V_{Phon} \sim V_{VT} + V_E \tag{6.3}$$

The representation of vocal tract variation (V_{VT}) as an analogue of genetic variance (V_G) is supported by the fact that vocal tract morphology exhibits heritable variation, with a portion of this variation being explained by genetic differences (Dediu et al., 2022). Chapter 2 demonstrated that this model can indeed be used to test hypotheses about language diversification, as it outlined that acoustic variation is proportional to vocal tract variation, and that inherent differences in speakers' vocal tracts act as a source of acoustic variation within speech communities.

One thing to keep it mind is that these proposed evolutionary models assume linear relationships between the phenotype and genotype (or, in this case, between

the phonotype and phenotype). It was shown in Chapter 2 that the linearity assumption is reasonable. A further consideration is the environmental variance (V_E), which may include factors such as diet, climate, social context, or cultural practices, all of which can influence the vocal characteristics of individuals or populations (Labov, 1966; Niedzielski, 1999; Maddieson, 2018; Blasi et al., 2019; Liang et al., 2023). In the *morpho-phonetic model of acoustic trait evolution*, V_E accounts for acoustic variation not attributable to phenotypic differences. Future studies should rigorously estimate V_E through longitudinal or experimental designs, similar to twin studies in genetics (Bouchard et al., 2013b; Dediu et al., 2022), but applied to acoustic traits (Przybyla et al., 1992; Nolan and Oh, 1996). This would allow for a more comprehensive model of acoustic trait evolution, integrating both genetic and environmental influences on phenotypic and phonotypic variation. Nevertheless, an alternative approach to examine environmental and cultural influences on language diversification is proposed below.

6.2.2 Neutral-like evolution as a null-model in acoustic diversification

The *morpho-phonetic model of acoustic trait evolution* established above enables testable hypotheses regarding evolutionary processes, such as distinguishing between neutral evolution and adaptive processes in acoustic trait diversification (Moran et al., 2024). Based on this model, phonotypic and phenotypic variation can be further divided into variation between (V_B) and within groups (V_W). As evolutionary theory predicts, under a neutral model (e.g., similar to genetic drift), variation between groups should be proportional to variation within groups (Lande, 1979; Lande, 1980), suggesting that differences between speech communities arise from stochastic processes that feed on the inherent within-community variation rather than selective pressures. In this case, the equation can be rewritten as:

$$V_B \sim V_W \tag{6.4}$$

This neutral-like model of acoustic diversification acts as a testable null-model in this case, i.e., a deviation from this proportionality, which can be tested using methods outlined in Chapters 2 and 4, would indicate non-random phonetic changes influenced by environmental or cultural factors[1]. This model allows us to identify instances where such factors (e.g., adaptive processes) override neutral modes of diversification. These models thus enable the disentangling of biological effects (e.g., vocal tract variation) from cultural or environmental influences on language diversification. Thus, it is proposed that the *morpho-phonetic model of acoustic trait evolution* and the null-model of neutral-like evolution should be applied more extensively in the language and speech sciences, for example, to investigate phonetic variation across human populations and to support integrative approaches in historical linguistics (Barbieri et al., 2022). Furthermore, inspired by recent studies demonstrating convergent vocal representations in the forebrain motor networks of birds and

[1]It is usually not possible to confirm that a trait evolves entirely without selection. Neutrality provides a useful baseline model representing the combined effects of many processes that may each be slightly non-neutral. Thus, the term "neutral-like" evolution is used here to denote evolutionary change that appears neutral but where true neutrality cannot be definitively established. Because distinguishing neutral evolution from stabilizing selection is often impractical, "neutral-like" is a more appropriate description.

humans (Bolhuis et al., 2010; Pfenning et al., 2014; Yang and Long, 2025), it is suggested that acoustic diversification models could be extended to other vocal learning species, such as birds (see Chapter 4), thereby providing potential insights into shared processes of vocal diversification across species.

Additionally, Chapters 4 and 5 expanded the neutral-like model to incorporate spatial effects, specifically isolation by distance in acoustic traits across populations (speech communities and populations of songbirds). Isolation by distance is a characteristic feature of neutral-like evolutionary processes, predicting that as populations disperse over time, genetic and phenotypic differences between the founding and descendant populations will increase. Any deviation from this expected pattern suggests the presence of non-neutral evolutionary influences (Lande, 1991; Slatkin and Excoffier, 2012). These findings also revisit longstanding questions about whether language evolution occurs gradually or in rapid, punctuated bursts (Atkinson et al., 2008). The evidence presented in this work suggests that different aspects of language (e.g., vocabulary vs. acoustic traits) may evolve at different tempos – that is, rates of change – and modes, meaning the patterns or mechanisms of evolution, such as gradual change versus punctuated bursts (Greenhill et al., 2017). This perspective aligns with the *mosaic approach to language evolution*, which views language as a complex system of distinct components, each evolving along different evolutionary paths (Parravicini and Pievani, 2018). These ideas further emphasize the need to integrate acoustic variation between individuals and speech communities into the study of language evolution.

Given that acoustic traits retain an evolutionary signal, potentially reflecting genealogical relationships between languages at the population-mean level (see Chapter 5), I propose that acoustic variation across languages may serve as a historical signal of linguistic divergence. Such acoustic data have been shown to reflect phylogenetic relationships in other vocal learning species, such as birds (Mason et al., 2017; Arato and Fitch, 2021; Rivera et al., 2023). Thus, acoustic data might be especially helpful for studying linguistic history where relationships between languages remain somewhat unclear (Endicott et al., 2003). The isolation by distance patterns discussed here also allow for testing how acoustic variation has spread globally. As an alternative to the highly debated Out-of-Africa hypothesis of phonemic diversity (Atkinson, 2011; Bybee, 2011; Atkinson, 2012; Cysouw et al., 2012; Tuyl and Pereltsvaig, 2012) – it is proposed here that these effects should be tested for acoustic variation across the world's languages.

In conclusion, this dissertation highlights how articulatory and acoustic variation between individuals – viewed as valuable information rather than being treated as noise – can serve as a rich source of information for models of speech and language evolution. By tracing how such variation emerges at the level of individual speakers and propagates across speaker communities, this study underscores its potential role in driving linguistic diversification. The findings gathered in this work point to the value of incorporating fine-grained articulatory data into evolutionary linguistics, both as empirical evidence and to inform theoretical frameworks. Moving forward, a deeper investigation into the morphological and anatomical influences on speech production is warranted to better understand how vocal tract structure shapes the emergence and distribution of sound systems across the world's languages. It is my hope that this research contributes to, and encourages, further interdisciplinary efforts in this direction.

Bibliography

Ackermann, Rebecca Rogers and James M Cheverud (2004). "Detecting genetic drift versus selection in human evolution". In: *Proceedings of the National Academy of Sciences* 101.52, pp. 17946–17951.

Adam, Iris et al. (2025). "A synchrotron X-ray CT-based 3D atlas of the songbird syrinx with single muscle fibre resolution implies fine motor control of syringeal vocal folds". In: *Philosophical Transactions B* 380.1920, p. 20230430.

Adams, Dean C (2016). "Evaluating modularity in morphometric data: Challenges with the RV coefficient and a new test measure". In: *Methods in Ecology and Evolution* 7.5, pp. 565–572.

Adams, Dean C, F James Rohlf, and Dennis E Slice (2004). "Geometric morphometrics: ten years of progress following the 'revolution'". In: *Italian journal of zoology* 71.1, pp. 5–16.

— (2013). "A field comes of age: geometric morphometrics in the 21st century". In: *Hystrix* 24.1, pp. 7–14.

Adams, Dean C. et al. (2022). *Geomorph: Software for geometric morphometric analyses. R package version 4.0.4*.

Adank, Patti, Roel Smits, and Roeland Van Hout (2004). "A comparison of vowel normalization procedures for language variation research". In: *The Journal of the Acoustical Society of America* 116.5, pp. 3099–3107.

Ahn, Emily and Eleanor Chodroff (2022). "Voxcommunis: A corpus for cross-linguistic phonetic analysis". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5286–5294.

Alatorre Warren, José Luis et al. (2019). "Evidence for independent brain and neurocranial reorganization during hominin evolution". In: *Proceedings of the National Academy of Sciences* 116.44, pp. 22115–22121.

Albessard-Ball, Lou and Antoine Balzeau (2018). "Of tongues and men: a review of morphological evidence for the evolution of language". In: *Journal of Language Evolution* 3.1, pp. 79–89.

Anderson, Cormac et al. (2023). "Variation in phoneme inventories: quantifying the problem and improving comparability". In: *Journal of Language Evolution* 8.2, pp. 149–168.

Anikin, Andrey, Rasmus Bååth, and Tomas Persson (2018). "Human non-linguistic vocal repertoire: call types and their meaning". In: *Journal of Nonverbal Behavior* 42.1, pp. 53–80.

Anthonissen, Lynn (2021). *Individuality in language change*. Vol. 360. Walter de Gruyter GmbH & Co KG.

Arato, Jozsef and W Tecumseh Fitch (2021). "Phylogenetic signal in the vocalizations of vocal learning and vocal non-learning birds". In: *Philosophical transactions of the royal society B* 376.1836, p. 20200241.

Ardila, Rosana et al. (2020). "Common Voice: A Massively-Multilingual Speech Corpus". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222.

Atkinson, Quentin D. (2011). "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion From Africa". In: *Science* 332, pp. 346–359.

— (2012). "Response to Comments on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa"". In: *Science* 335.6069, pp. 657–657.

Atkinson, Quentin D and Russell D Gray (2005). "Curious parallels and curious connections – phylogenetic thinking in biology and historical linguistics". In: *Systematic Biology* 54.4, pp. 513–526.

Atkinson, Quentin D et al. (2008). "Languages evolve in punctuational bursts". In: *Science* 319.5863, pp. 588–588.

Balezeau, Fabien et al. (2020). "Primate auditory prototype in the evolution of the arcuate fasciculus". In: *Nature neuroscience* 23.5, pp. 611–614.

Barbieri, Chiara et al. (2022). "A global analysis of matches and mismatches between human genetic and linguistic histories". In: *Proceedings of the National Academy of Sciences* 119.47, e2122084119.

Barham, Lawrence and Daniel Everett (2021). "Semiotics and the origin of language in the Lower Palaeolithic". In: *Journal of Archaeological Method and Theory* 28.2, pp. 535–579.

Bartoli, Eleonora et al. (2015). "Listener–speaker perceived distance predicts the degree of motor contribution to speech perception". In: *Cerebral Cortex* 25.2, pp. 281–288.

Beddor, Patrice Speeter (2009). "A coarticulatory path to sound change". In: *Language* 85.4, pp. 785–821.

Bell, Ben D et al. (2004). "Short records of marsh warbler (*Acrocephalus palustris*) song provide indices that correlate with nesting success". In: *Journal of Ornithology* 145, pp. 8–15.

Benítez-Burraco, Antonio and Steven Moran, eds. (2018). *The Adaptive Value of Languages: Non-linguistic Causes of Language Diversity*. Frontiers in Psychology.

— eds. (2024). *The Adaptive Value of Languages: Non-linguistic Causes of Language Diversity, Volume II*. Frontiers in Psychology.

Bentz, Christian et al. (2018). "The evolution of language families is shaped by the environment beyond neutral drift". In: *Nature Human Behaviour* 2.11, pp. 816–821.

Berwick, Robert C and Noam Chomsky (2016). *Why only us: Language and evolution*. MIT press.

Birkholz, Peter et al. (2020). "Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties". In: *Scientific data* 7.1, p. 255.

Bladon, R.A.W. (1982). "Arguments against formants in the auditory representation of speech". In: *The Representation of Speech in the Peripheral Auditory System*. Ed. by R Calson and B Granström. Elsevier Biomedical, Amsterdam, pp. 95–102.

Blasi, Damián E et al. (2019). "Human sound systems are shaped by post-Neolithic changes in bite configuration". In: *Science* 363.6432, eaav3218.

Blumstein, Sheila E and Kenneth N Stevens (1979). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants". In: *The Journal of the Acoustical Society of America* 66.4, pp. 1001–1017.

Blythe, Richard A (2015). "Colloquium: Hierarchy of scales in language dynamics". In: *The European Physical Journal B* 88.11, p. 295.

Boë, Louis-Jean et al. (2002). "The potential Neandertal vowel space was as large as that of modern humans". In: *Journal of Phonetics* 30.3, pp. 465–484.

Boë, Louis-Jean et al. (2019). "Which way to the dawn of speech?: Reanalyzing half a century of debates and data in light of speech science". In: *Science Advances* 5.12, eaaw3916.

Boersma, Paul (2001). "Praat, a system for doing phonetics by computer". In: *Glot. Int.* 5.9, pp. 341–345.

— (2014). "Acoustic analysis". In: *Research Methods in Linguistics*. Ed. by Robert J. Podesva and Devyani Sharma. Cambridge University Press, 375–396.

Bolhuis, Johan J, Kazuo Okanoya, and Constance Scharff (2010). "Twitter evolution: converging mechanisms in birdsong and human speech". In: *Nature Reviews Neuroscience* 11.11, pp. 747–759.

Bolhuis, Johan J et al. (2014). "How could language have evolved?" In: *PLOS biology* 12.8, e1001934.

Bolles, Edmund Blair (2011). "Speech and Birdsongs Reflect Convergent Evolutionary History". In: *BioScience* 61.3, pp. 248–248.

Boncoraglio, Giuseppe and Nicola Saino (2007). "Habitat structure and the evolution of bird song: a meta-analysis of the evidence for the acoustic adaptation hypothesis". In: *Functional Ecology* 21, pp. 134–142.

Bookstein, Fred L (1996). "Applying landmark methods to biological outline data". In: *Image fusion and shape variability techniques*. Leeds University Press.

— (2018). *A Course in Morphometrics for Biologists: Geometry and Statistics for Studies of Organismal Form*. Cambridge University Press.

Bookstein, Fred L and Philipp Mitteroecker (2014). "Comparing covariance matrices by relative eigenanalysis, with applications to organismal biology". In: *Evolutionary biology* 41, pp. 336–350.

Bookstein, Fred L et al. (2003). "Cranial integration in Homo: singular warps analysis of the midsagittal plane in ontogeny and evolution". In: *Journal of Human Evolution* 44.2, pp. 167–187.

Bouchard, Kristofer E et al. (2013a). "Functional organization of human sensorimotor cortex for speech articulation". In: *Nature* 495.7441, pp. 327–332.

Bouchard, Kristofer E et al. (2016). "High-resolution, non-invasive imaging of upper vocal tract articulators compatible with human brain recordings". In: *PLOS One* 11.3, e0151327.

Bouchard, Thomas J et al. (2013b). "Sources of human psychological differences: The Minnesota study of twins reared apart". In: *Personality and Personality Disorders*. Routledge, pp. 139–144.

Bouchard-Côté, Alexandre et al. (2013). "Automated reconstruction of ancient languages using probabilistic models of sound change". In: *Proceedings of the National Academy of Sciences* 110.11, pp. 4224–4229.

Bresch, Erik et al. (2006). "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans". In: *The Journal of the Acoustical Society of America* 120.4, pp. 1791–1794.

Brochhagen, Thomas et al. (2023). "From language development to language evolution: A unified view of human lexical creativity". In: *Science* 381.6656, pp. 431–436.

Bromham, Lindell, Keaghan J Yaxley, and Marcel Cardillo (2024). "Islands are engines of language diversity". In: *Nature Ecology & Evolution* 8.10, pp. 1991–2002.

Browman, Catherine P and Louis Goldstein (1992). "Articulatory phonology: An overview". In: *Phonetica* 49.3-4, pp. 155–180.

Bryant, David and Vincent Moulton (2004). "Neighbor-net: an agglomerative method for the construction of phylogenetic networks". In: *Molecular biology and evolution* 21.2, pp. 255–265.

Budd, Graham E (2021). "Morphospace". In: *Current Biology* 31.19, R1181–R1185.
Burridge, James (2023). "Vowel Normalisation in Latent Space for Sociolinguistics". In: *Proceedings of INTERSPEECH 2023*. Dublin, Ireland, pp. 3547–3551.
Bybee, Joan (2011). "How plausible is the hypothesis that population size and dispersal are related to phoneme inventory size? Introducing and commenting on a debate". In: *Linguistic Typology* 15.2.
Campbell, Lyle (2013). *Historical linguistics*. Edinburgh University Press.
Campbell, Lyle and William J Poser (2008). *Language classification*. Cambridge University Press.
Catford, John Cunnison (1977). *Fundamental problems in phonetics*. Bloomington: Indiana University Press.
Cavalcanti, Julio Cesar, Anders Eriksson, and Plinio A Barbosa (2021). "Acoustic analysis of vowel formant frequencies in genetically-related and non-genetically related speakers with implications for forensic speaker comparison". In: *PLOS One* 16.2, e0246645.
Cavalli-Sforza, Luigi Luca, Paolo Menozzi, and Alberto Piazza (1994). *The history and geography of human genes*. Princeton university press.
Chang, Edward F et al. (2010). "Categorical speech representation in human superior temporal gyrus". In: *Nature Neuroscience* 13.11, pp. 1428–1432.
Chang, Yi-Chung et al. (2014). "Synchronized imaging and acoustic analysis of the upper airway in patients with sleep-disordered breathing". In: *Physiol. Meas.* 35.12, p. 2501.
Chapuis, C (1971). "Un exemple de l'influence du milieu sur les émissions vocales des oiseaux: l'évolution des chants en forêt équatoriale". In: *Revue d'Écologie* 2, pp. 183–202.
Chartier, Josh et al. (2018). "Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex". In: *Neuron* 98.5, pp. 1042–1054.
Cheung, Connie et al. (2016). "The auditory representation of speech sounds in human motor cortex". In: *eLife* 5, e12577.
Chomsky, Noam (2005). *Rules and representations*. Columbia University Press.
Christiansen, Morten H and Nick Chater (2008). "Language as shaped by the brain". In: *Behavioral and brain sciences* 31.5, pp. 489–509.
Christiansen, Morten H and Simon Kirby (2003a). *Language evolution*. Oxford University Press.
Christiansen, Morten H. and Simon Kirby (2003b). "Language evolution: Consensus and controversies". In: *Trends in Cognitive Sciences* 7.7, pp. 300–307.
Chung, Hyunju et al. (2012). "Cross-linguistic studies of children's and adults' vowel spaces". In: *The Journal of the Acoustical Society of America* 131.1, pp. 442–454.
Clopper, Cynthia G, David B Pisoni, and Kenneth De Jong (2005). "Acoustic characteristics of the vowel systems of six regional varieties of American English". In: *The Journal of the Acoustical society of America* 118.3, pp. 1661–1676.
Collyer, M. L. and D. C. Adams (2018). "RRPP: An R package for fitting linear models to high-dimensional data using residual randomization." In: *Methods in Ecology and Evolution* 9 (2), pp. 1772–1779.
Conde-Valverde, Mercedes et al. (2021). "Neanderthals and Homo sapiens had similar auditory and speech capacities". In: *Nature ecology & evolution* 5.5, pp. 609–615.
Creanza, Nicole et al. (2015). "A comparison of worldwide phonemic and genetic variation in human populations". In: *Proceedings of the National Academy of Sciences* 112.5, pp. 1265–1272.

Crelin, Edmund S (1987). *The human vocal tract: Anatomy, function, development, and evolution*. Vantage Press.

Cysouw, Michael, Dan Dediu, and Steven Moran (2012). "Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa"". In: *Science* 335.6069, pp. 657–657.

Darwin, Charles (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Davis, Steven and Paul Mermelstein (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: *IEEE transactions on acoustics, speech, and signal processing* 28.4, pp. 357–366.

De Boer, Bart (2000). "Self-organization in vowel systems". In: *Journal of phonetics* 28.4, pp. 441–465.

— (2001). *The origins of vowel systems*. Vol. 1. Oxford University Press, USA.

De Boer, Bart and W Tecumseh Fitch (2010). "Computer models of vocal tract evolution: An overview and critique". In: *Adaptive Behavior* 18.1, pp. 36–47.

De Framond, Léna and Henrik Brumm (2022). "Long-term effects of noise pollution on the avian dawn chorus: a natural experiment facilitated by the closure of an international airport". In: *Proceedings of the Royal Society B* 289.1982, p. 20220906.

Dediu, Dan, Rick Janssen, and Scott R. Moisik (2017). "Language Is Not Isolated From Its Wider Environment: Vocal Tract Influences on the Evolution of Speech and Language". In: *Language & Communication*, pp. 9–20.

Dediu, Dan, Rick Janssen, and Scott R Moisik (2019). "Weak biases emerging from vocal tract anatomy shape the repeated transmission of vowels". In: *Nature Human Behaviour* 3.10, pp. 1107–1115.

Dediu, Dan and D Robert Ladd (2007). "Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, *ASPM* and *Microcephalin*". In: *Proceedings of the National Academy of Sciences* 104.26, pp. 10944–10949.

Dediu, Dan and Stephen C. Levinson (2013). "On the Antiquity of Language: The Reinterpretation of Neandertal Linguistic Capacities and Its Consequences". In: *Frontiers in Psychology* 4, pp. 1–17.

Dediu, Dan and Stephen C Levinson (2018). "Neanderthal language revisited: not only us". In: *Current Opinion in Behavioral Sciences* 21, pp. 49–55.

Dediu, Dan and Scott R Moisik (2019). "Pushes and pulls from below: Anatomical variation, articulation and sound change". In: *Glossa: a journal of general linguistics*.

Dediu, Dan et al. (2021). "The vocal tract as a time machine: Inferences about past speech and language from the anatomy of the speech organs". In: *Philosophical Transactions of the Royal Society B* 376.1824, p. 20200192.

Dediu, Dan et al. (2022). "The heritability of vocal tract structures estimated from structural MRI in a large cohort of Dutch twins". In: *Human genetics* 141.12, pp. 1905–1923.

Disner, Sandra Ferrari (1980). "Evaluation of vowel normalization procedures". In: *Journal of the Acoustical Society of America* 67.1, pp. 253–261.

Donohue, Mark and Johanna Nichols (2011). "Does Phoneme Inventory Size Correlate With Population Size?" In: *Linguistic Typology* 15.2, pp. 161–170.

Doucette, Amanda et al. (2024). "Investigating the universality of consonant and vowel co-occurrence restrictions". In: *Glossa: a journal of general linguistics* 9.1.

Doupe, Allison J and Patricia K Kuhl (1999). "Birdsong and human speech: common themes and mechanisms". In: *Annual Review of Neuroscience* 22.1, pp. 567–631.

Dowsett-Lemaire, Françoise (1979). "The imitative range of the song of the marsh warbler *Acrocephalus palustris*, with special reference to imitations of African birds". In: *Ibis* 121.4, pp. 453–468.

Dryden, Ian L and Kanti V Mardia (2016). *Statistical Shape Analysis: With Applications in R*. Vol. 995. John Wiley & Sons.

Dubreuil, Benoît and Christopher Stuart Henshilwood (2013). "Archeology and the language-ready brain". In: *Language and Cognition* 5.2-3, pp. 251–260.

Duckworth, Martin et al. (2011). "Improving the consistency of formant measurement". In: *The International Journal of Speech, Language and the Law* 18.1, pp. 35–51.

Ekström, Axel G (2022). "Motor constellation theory: A model of infants' phonological development". In: *Frontiers in Psychology* 13.

Ellis, Daniel P. W. (2005). *PLP and RASTA (and MFCC, and inversion) in Matlab*. Accessed on 05.19., 2023.

Endicott, Phillip et al. (2003). "The genetic origins of the Andaman Islanders". In: *The American Journal of Human Genetics* 72.1, pp. 178–184.

Erattakulangara, Subin et al. (2025). "Open-Source Manually Annotated Vocal Tract Database for Automatic Segmentation from 3D MRI Using Deep Learning: Benchmarking 2D and 3D Convolutional and Transformer Networks". In: *arXiv preprint arXiv:2501.06229*.

Everett, Caleb (2013). "Evidence for Direct Geographic Influences on Linguistic Sounds: The Case of Ejectives". In: *PLOS One* 8.6, e65275.

Everett, Caleb, Damián E. Blasi, and Seán G. Roberts (2015). "Climate, Vocal Folds, and Tonal Languages: Connecting the Physiological and Geographic Dots". In: *Proceedings of the National Academy of Sciences* 112.5, pp. 1322–1327.

Falconer, Douglas Scott (1996). *Introduction to quantitative genetics*. Pearson Education India.

Falzone, Alessandra, Alessandra Anastasi, Antonio Pennisi, et al. (2014). "Biological constraints and evolution of language: a hyphotesis on the exaptation of human vocal structures". In: *Proceedings of the European Society for the Study of Human Evolution*, pp. 66–67.

Fant, Gunnar (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Figueroa, Mauricio and Young Shin Kim (2021). "Measuring Consonants". In: *The Cambridge Handbook of Phonetics*. Ed. by Rachael-Anne Knight and Jane Setter. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 285–311.

Fitch, W Tecumseh (1997). "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques". In: *The Journal of the Acoustical Society of America* 102.2, pp. 1213–1222.

— (2000). "The evolution of speech: a comparative review". In: *Trends in cognitive sciences* 4.7, pp. 258–267.

— (2006). "Production of vocalizations in mammals". In: *Visual Communication* 3.145, pp. 115–121.

— (2010). *The Evolution of Language*. Cambridge University Press.

Fitch, W Tecumseh and Jay Giedd (1999). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging". In: *The Journal of the Acoustical Society of America* 106.3, pp. 1511–1522.

Fitch, W Tecumseh and David Reby (2001). "The descended larynx is not uniquely human". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268.1477, pp. 1669–1675.

Fitch, W Tecumseh et al. (2016). "Monkey vocal tracts are speech-ready". In: *Science advances* 2.12, e1600723.

Fox, Anthony (1995). *Linguistic reconstruction: An introduction to theory and method*. Oxford University Press, USA.

Friedrichs, Daniel et al. (2017). "Vowel recognition at fundamental frequencies up to 1 kHz reveals point vowels as acoustic landmarks". In: *The Journal of the Acoustical Society of America* 142.2, pp. 1025–1033.

Fry, Dennis Butler (1979). *The physics of speech*. Cambridge University Press.

Fu, Maojing et al. (2017). "High-frame-rate full-vocal-tract 3D dynamic speech imaging". In: *Magnetic resonance in medicine* 77.4, pp. 1619–1629.

Fuchs, Susanne, Ralf Winkler, and Pascal Perrier (2008). "Do speakers' vocal tract geometries shape their articulatory vowel space?" In: *ISSP 2008-8th International Seminar on Speech Production*, pp. 333–336.

Gaines, Jessica L et al. (2021). "Discrete constriction locations describe a comprehensive range of vocal tract shapes in the Maeda model". In: *JASA Express Letters* 1.12, p. 124402.

Galantucci, Bruno, Carol A Fowler, and Michael T Turvey (2006). "The motor theory of speech perception reviewed". In: *Psychonomic bulletin & review* 13.3, pp. 361–377.

Garcia, Maxime et al. (2017). "Acoustic allometry revisited: Morphological determinants of fundamental frequency in primate vocal production". In: *Scientific Reports* 7.1, p. 10450.

Geng, Christian and Christine Mooshammer (2009). "How to stretch and shrink vowel systems: Results from a vowel normalization procedure". In: *Journal of the Acoustical Society of America* 125.5, pp. 3278–3288.

Ghazanfar, Asif A and Drew Rendall (2008). "Evolution of human vocal production". In: *Current Biology* 18.11, R457–R460.

Gick, Bryan, Ian Wilson, and Donald Derrick (2013). *Articulatory Phonetics*. United States: Wiley-Blackwell.

Gick, Bryan et al. (2019). "Articulatory phonetics". In: *The Routledge Handbook of Phonetics*. Routledge, pp. 107–125.

Gisladottir, Rosa S et al. (2023). "Sequence variants affecting voice pitch in humans". In: *Science Advances* 9.23, eabq2969.

Giuliani, Diego, Matteo Gerosa, and Fabio Brugnara (2006). "Improved automatic speech recognition through speaker normalization". In: *Computer Speech & Language* 20.1, pp. 107–123.

Gower, John C (1975). "Generalized Procrustes analysis". In: *Psychometrika* 40.1, pp. 33–51.

Gray, Russell (2005). "Pushing the time barrier in the quest for language roots". In: *Science* 309.5743, pp. 2007–2008.

Gray, Russell D. and Quentin D. Atkinson (2003). "Language-tree divergence times support the Anatolian theory of Indo-European origin". In: *Nature*, pp. 435–439.

Gray, Russell D. and Fiona M. Jordan (2000). "Language trees support the express-train sequence of Austronesian expansion". In: *Nature* 405, pp. 1052–1055.

Greenhill, Simon J. (2014). "Demographic Correlates of Language Diversity". In: *The Routledge Handbook of Historical Linguistics*. Ed. by Claire Bowern and Bethwyn Evans. Routledge, pp. 557–578.

— (2023). "A shared foundation of language change". In: *Science* 381.6656, pp. 374–375.

Greenhill, Simon J et al. (2017). "Evolutionary dynamics of language systems". In: *Proceedings of the National Academy of Sciences* 114.42, E8822–E8829.

Guenther, Frank H (2016). *Neural Control of Speech*. MIT Press.
Gully, Amelia J. (2021). "Quantifying Vocal Tract Shape Variation and its Acoustic Impact: A Geometric Morphometric Approach". In: *Proc. Interspeech 2021*, pp. 3999–4003.
Gunz, Philipp, Philipp Mitteroecker, and Fred L Bookstein (2005). "Semilandmarks in three dimensions". In: *Modern morphometrics in physical anthropology*. Springer, pp. 73–98.
Hammarström, Harald et al. (2024). *Glottolog 5.1*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Harrington, Jonathan, Steve, and Cassidy (1999). *Techniques in speech acoustics*. Springer Science & Business Media.
Harrison, Philip Thomas (2013). "Making Accurate Formant Measurements: An Empirical Investigation of the Influence of the Measurement Tool, Analysis Settings and Speaker on Formant Measurements". PhD thesis. University of York.
Haudricourt, André G. (1961). "Richesse en phonèmes et richesse en locuteurs". In: *L'Homme* 1, pp. 5–10.
Hauser, Marc D, Noam Chomsky, and W Tecumseh Fitch (2002). "The faculty of language: what is it, who has it, and how did it evolve?" In: *Science* 298.5598, pp. 1569–1579.
Hauser, Marc D et al. (2014). "The mystery of language evolution". In: *Frontiers in psychology* 5, p. 88824.
Hay, Jennifer and Laurie Bauer (2007). "Phoneme Inventory Size and Population Size". In: *Language* 83, pp. 388–400.
Heggarty, Paul et al. (2023). "Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages". In: *Science* 381.6656, eabg0818.
Hermansky, Hynek, Jordan R Cohen, and Richard M Stern (2013). "Perceptual properties of current speech recognition technology". In: *Proceedings of the IEEE* 101.9, pp. 1968–1985.
Hijmans, Robert J. (2024). *geosphere: Spherical Trigonometry*. R package version 1.5-20.
Holden, Clare Janaki (2002). "Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269.1493, pp. 793–799.
Holmes, John N, Wendy J Holmes, and Philip N Garner (1997). "Using formant frequencies in speech recognition." In: *Eurospeech*. Vol. 97, pp. 2083–2087.
Horvatić, Sven et al. (2021). "Correlation between acoustic divergence and phylogenetic distance in soniferous European gobiids (Gobiidae; Gobius lineage)". In: *PLOS one* 16.12, e0260810.
Hunley, Keith et al. (2008). "Genetic and linguistic coevolution in Northern Island Melanesia". In: *PLOS genetics* 4.10, e1000239.
Hurford, James R. (2011). *The Origins of Grammar: Language in the Light of Evolution*. Oxford University Press.
Hyland Bruno, Julia et al. (2021). "Birdsong learning and culture: analogies with human spoken language". In: *Annual Review of Linguistics* 7.1, pp. 449–472.
Iivonen, Antti (1994). "A psychoacoustical explanation for the number of major IPA vowels". In: *Journal of the International Phonetic Association* 24.2, 73–90.
Il'ina, I Yu et al. (2021). "Song divergence in the group of "Brownish" Chiffchaff forms (*Phylloscopus collybita* sensu lato)". In: *Biology Bulletin* 48, pp. 1067–1074.
International Phonetic Association (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Isaieva, Karyna et al. (2023). "Super-resolved dynamic 3D reconstruction of the vocal tract during natural speech". In: *Journal of Imaging* 9.10, p. 233.

Ito, Masashi, Jun Tsuchida, and Masafumi Yano (2001). "On the effectiveness of whole spectral shape for vowel perception". In: *The Journal of the Acoustical Society of America* 110.2, pp. 1141–1149.

J. Sueur, T. Aubin, and C. Simonis (2008). "Seewave: a free modular tool for sound analysis and synthesis". In: *Bioacoustics* 18, pp. 213–226.

James, Logan S et al. (2021). "Phylogeny and mechanisms of shared hierarchical patterns in birdsong". In: *Current Biology* 31.13, pp. 2796–2808.

Janson, Tore (1986). "Cross-linguistic trends in the frequency of CV sequences". In: *Phonology* 3, pp. 179–195.

Janssen, Rick, Scott R Moisik, and Dan Dediu (2019). "The effects of larynx height on vowel production are mitigated by the active control of articulators". In: *Journal of Phonetics* 74, pp. 1–17.

Johansson, Sverker (2015). "Language abilities in Neanderthals". In: *Annu. Rev. Linguist.* 1.1, pp. 311–332.

Johnson, Keith (2023). "Individual differences in speech production: What is "phonetic substance?"" In: *20th International Congress of Phonetic Sciences (ICPhS)*, pp. 1126–1130.

Johnson, Keith, Peter Ladefoged, and Mona Lindau (1993). "Individual differences in vowel production". In: *The Journal of the Acoustical Society of America* 94.2, pp. 701–714.

Johnson, Keith and Matthias J Sjerps (2021). "Speaker normalization in speech perception". In: *The Handbook of Speech Perception*. Ed. by Jennifer S. Pardo et al. Wiley Online Library, pp. 145–176.

Josserand, Mathilde et al. (2024). "Adapting to individual differences: An experimental study of language evolution in heterogeneous populations". In: *Cognitive Science* 48.11, e70011.

Kawasaki, Haruko and John J Ohala (1980). "Acoustic basis for universal constraints on sound sequences". In: *The Journal of the Acoustical Society of America* 68.S1, S33–S33.

Kendall, David G (1981). "The statistics of shape". In: *Interpreting multivariate data*. Ed. by V Barnett. New York: John Wiley & Sons New York, pp. 75–80.

— (1984). "Shape manifolds, procrustean metrics, and complex projective spaces". In: *Bull. Lond. Math. Soc.* 16.2, pp. 81–121.

— (1985). "Exact distributions for shapes of random triangles in convex sets". In: *Adv. Appl. Probab.* 17.2, pp. 308–329.

Kent, Ray D and Carrie Rountrey (2020). "What acoustic studies tell us about vowels in developing and disordered speech". In: *American Journal of Speech-Language Pathology* 29.3, pp. 1749–1778.

Khanna, Arjun R et al. (2024). "Single-neuronal elements of speech production in humans". In: *Nature*, pp. 1–8.

Kiefte, Michael, Terrance M Nearey, and Peter F Assmann (2013). "Vowel perception in normal speakers". In: *Handbook of vowels and vowel disorders*. Psychology Press, pp. 160–185.

Kirby, Simon, Hannah Cornish, and Kenny Smith (2008). "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language". In: *Proceedings of the National Academy of Sciences* 105.31, pp. 10681–10686.

Klein, Wolfgang, Reinier Plomp, Louis CW Pols, et al. (1970). "Vowel spectra, vowel spaces, and vowel identification". In: *Journal of the Acoustical Society of America* 48.4, pp. 999–1009.

Klingenberg, Christian Peter (2015). "Analyzing fluctuating asymmetry with geometric morphometrics: concepts, methods, and applications". In: *Symmetry* 7.2, pp. 843–934.

— (2016). "Size, shape, and form: concepts of allometry in geometric morphometrics". In: *Development genes and evolution* 226.3, pp. 113–137.

Koile, Ezequiel et al. (2022). "Geography and language divergence: The case of Andic languages". In: *PLOS One* 17.5, e0265460.

Konopka, Genevieve et al. (2009). "Human-specific transcriptional regulation of CNS development genes by FOXP2". In: *Nature* 462.7270, pp. 213–217.

Krause, Bernard L (1987). "Bioacoustics, habitat ambience in ecological balance". In: *Whole Earth Review* 57.472, pp. 14–18.

— (1993). "The niche hypothesis". In: *Soundsc Newsl* 6, p. 6.

Kreiman, Jody and Diana Sidtis (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

Labov, William (1966). *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

— (2001). *Principles of Linguistic Change. Volume 2: Social Factors*. Blackwell.

Lachlan, RF and MR Servedio (2004). "Song learning accelerates allopatric speciation". In: *Evolution* 58.9, pp. 2049–2063.

Lachlan, Robert F. (2008). "The Evolution of Flexibility in Bird Song". In: *Evolution of Communicative Flexibility: Complexity, Creativity, and Adaptability in Human and Animal Communication*. The MIT Press.

Lachlan, Robert F et al. (2013). "The progressive loss of syntactical structure in bird song along an island colonization chain". In: *Current Biology* 23.19, pp. 1896–1901.

Lachlan, Robert Francis (2000). "Cultural evolution in bird song: in theory and in chaffinches (Fringilla coelebs)". PhD thesis. University of St Andrews.

Ladefoged, Peter (1996). *Elements of acoustic phonetics*. University of Chicago Press.

Ladefoged, Peter and Donald Eric Broadbent (1957). "Information conveyed by vowels". In: *The Journal of the Acoustical Society of America* 29.1, pp. 98–104.

Ladefoged, Peter and Keith Johnson (2014). *A Course in Phonetics*. Cengage Learning.

Ladefoged, Peter and Ian Maddieson (1996). *The sounds of the world's languages*. Vol. 1012. Blackwell Oxford.

Ladefoged, Peter et al. (1972). "An auditory-motor theory of speech production". In: *UCLA Working Papers in Phonetics* 22.48, pp. 48–76.

Ladoukakis, Emmanuel D, Dimitris Michelioudakis, and Elena Anagnostopoulou (2022). "Toward an evolutionary framework for language variation and change". In: *BioEssays* 44, p. 2100216.

Lagler, KF, JE Bardach, and RR Miller (1962). *Ichthyology*. New York: John Wiley & Sons.

Lammert, Adam, Michael Proctor, and Shrikanth Narayanan (2013). "Interspeaker variability in hard palate morphology and vowel production". In: *Journal of Speech, Language, and Hearing Research* 56, S1924–S1933.

Lammert, Adam et al. (2011). "Morphological variation in the adult vocal tract: A modeling study of its potential acoustic impact". In: *Twelfth Annual Conference of the International Speech Communication Association*, pp. 2813–2816.

Lande, R (1991). "Isolation by distance in a quantitative trait." In: *Genetics* 128.2, pp. 443–452.

Lande, Russell (1979). "Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry". In: *Evolution*, pp. 402–416.
— (1980). "Genetic variation and phenotypic evolution during allopatric speciation". In: *The American Naturalist* 116.4, pp. 463–479.
Lande, Russell and Stevan J Arnold (1983). "The measurement of selection on correlated characters". In: *Evolution*, pp. 1210–1226.
Landi, Federica and Paul O'Higgins (2019). "Applying geometric morphometrics to digital reconstruction and anatomical investigation". In: *Biomedical Visualisation: Volume 4*, pp. 55–71.
Lane, Harlan and Bernard Tranel (1971). "The Lombard sign and the role of hearing in speech". In: *Journal of Speech and Hearing Research* 14.4, pp. 677–709.
Lautenschlager, Stephan (2016). "Digital reconstruction of soft-tissue structures in fossils". In: *The Paleontological Society Papers* 22, pp. 101–117.
Lavan, Nadine and Carolyn McGettigan (2023). "A model for person perception from familiar and unfamiliar voices". In: *Communications Psychology* 1.1, p. 1.
Laver, John (1980). *The phonetic description of voice quality*. Cambridge University Press.
Le Maître, Anne and Philipp Mitteroecker (2019). "Multivariate comparison of variance in R". In: *Methods in Ecology and Evolution* 10.9, pp. 1380–1392.
Leonard, Matthew K et al. (2023). "Large-scale single-neuron speech sound encoding across the depth of human cortex". In: *Nature*, pp. 1–10.
Levinson, Stephen C and Dan Dediu (2013). "The interplay of genetic and cultural factors in ongoing language evolution". In: *Cultural evolution: Society, technology, language, and religion. Strüngmann Forum Reports*. Vol. 12. MIT Press, pp. 219–232.
Lewontin, Richard C (1970). "The units of selection". In: *Annual review of ecology and systematics*, pp. 1–18.
Liang, Yuzhu et al. (2023). "Languages in China link climate, voice quality, and tone in a causal chain". In: *Humanities and Social Sciences Communications* 10.453, pp. 1–10.
Liberman, Alvin M and Ignatius G Mattingly (1985). "The motor theory of speech perception revised". In: *Cognition* 21.1, pp. 1–36.
Lieberman, Philip (2007). "The evolution of human speech: Its anatomical and neural bases". In: *Current Anthropology* 48.1, pp. 39–66.
— (2015). "Language did not spring forth 100,000 years ago". In: *PLOS biology* 13.2, e1002064.
Lieberman, Philip and Sheila E Blumstein (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge University Press.
Lieberman, Philip and Edmund S Crelin (1972). "On the Speech of Neanderthal Man". In: *The Speech of Primates* 148, p. 76.
Lieberman, Philip, Edmund S Crelin, and Dennis H Klatt (1972). "Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee". In: *American anthropologist* 74.3, pp. 287–307.
Lieberman, Philip H, Dennis H Klatt, and William H Wilson (1969). "Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates". In: *Science* 164.3884, pp. 1185–1187.
Liljencrants, Johan and Björn Lindblom (1972). "Numerical simulation of vowel quality systems: The role of perceptual contrast". In: *Language*, pp. 839–862.
Lim, Yongwan et al. (2019). "3D dynamic MRI of the vocal tract during natural speech". In: *Magnetic resonance in medicine* 81.3, pp. 1511–1520.

Lim, Yongwan et al. (2021). "A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images". In: *Scientific Data* 8.1, pp. 1–14.

Lin, Susan (2021). "Observing and Measuring Speech Articulation". In: *The Cambridge Handbook of Phonetics*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 362—386.

Lindblom, Björn (1963). "Spectrographic study of vowel reduction". In: *The Journal of the Acoustical Society of America* 35.11, pp. 1773–1781.

— (1990a). "Explaining phonetic variation: A sketch of the H&H theory". In: *Speech Production and Speech Modelling*. Ed. by W.J. Hardcastle and A. Marchal. Springer, pp. 403–439.

— (1990b). "Models of phonetic variation and selection". In: *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm* 11, pp. 65–100.

— (1990c). "On the notion of "possible speech sound"". In: *Journal of phonetics* 18.2, pp. 135–152.

Lindblom, Björn and Peter MacNeilage (2011). "Coarticulation: A universal phonetic phenomenon with roots in deep time". In: *TMH-QPSR (Speech, Music and Hearing Quarterly Progress and Status Report)* 51, pp. 41–44.

Lindblom, Björn EF and Johan EF Sundberg (1971). "Acoustical consequences of lip, tongue, jaw, and larynx movement". In: *The Journal of the Acoustical Society of America* 50.4B, pp. 1166–1179.

List, Johann-Mattis et al. (2014). "Networks of lexical borrowing and lateral gene transfer in language and genome evolution". In: *Bioessays* 36.2, pp. 141–150.

Lobanov, B. M. (1971). "Classification of Russian vowels spoken by different speakers". In: *The Journal of the Acoustical Society of America* 49.2, pp. 606–608.

Logan, Beth (2000). "Mel frequency cepstral coefficients for music modeling." In: *International Symposium on Music Information Retrieval*. 1, pp. 1–11.

Lombard, Ètienne (1911). "Le signe de l'élévation de la voix [The sign of voice rise]". In: *Annales des Maladies de l'Oreille et du Larynx* 37.2, pp. 101–119.

Longtine, Charlie et al. (2024). "Homology and the evolution of vocal folds in the novel avian voice box". In: *Current Biology* 34.3, pp. 461–472.

Lynch, Alejandro (1996). "The population memetics of birdsong". In: *Ecology and Evolution of Acoustic Communication in Birds*, pp. 181–197.

Lynch, Alejandro and Allan J Baker (1993). "A population memetics approach to cultural evolution in chaffinch song: meme diversity within populations". In: *The American Naturalist* 141.4, pp. 597–620.

MacKenzie, Laurel (2019). "Perturbing the community grammar: Individual differences and community-level constraints on sociolinguistic variation". In: *Glossa: a journal of general linguistics* 4.1.

MacNeilage, Peter F and Barbara L Davis (2000). "On the origin of internal structure of word forms". In: *Science* 288.5465, pp. 527–531.

Maddieson, Ian (2018). "Language adapts to environment: Sonority and temperature". In: *Frontiers in Communication* 3, p. 28.

Maddieson, Ian and Karl Benedict (2023). "Demonstrating environmental impacts on the sound structure of languages: Challenges and solutions". In: *Frontiers in Psychology* 14, p. 1200463.

Maddieson, Ian and Coupé (2015). "Human Spoken Language Diversity and the Acoustic Adaptation Hypothesis". In: *The Journal of the Acoustical Society of America* 138.3, pp. 1838–1838.

Maddieson, Ian and Kristin Precoda (1992). "Syllable structure and phonetic models". In: *Phonology*, pp. 45–60.

Marcus, Leslie F (1990). "Traditional morphometrics". In: *Proceedings of the Michigan morphometrics workshop*. Vol. 2. The University of Michigan Museum of Zoology Ann Arbor, USA, pp. 77–122.

Maricic, Tomislav et al. (2013). "A recent evolutionary change affects a regulatory element in the human FOXP2 gene". In: *Molecular Biology and Evolution* 30.4, pp. 844–852.

Mason, Nicholas A et al. (2017). "Song evolution, speciation, and vocal learning in passerine birds". In: *Evolution* 71.3, pp. 786–796.

Maurer, Dieter (2016). *Acoustics of the vowel-preliminaries*. Peter Lang International Academic Publishers.

McMurray, Bob (2022). "The myth of categorical perception". In: *The Journal of the Acoustical Society of America* 152.6, pp. 3819–3842.

Mesgarani, Nima et al. (2014). "Phonetic feature encoding in human superior temporal gyrus". In: *Science* 343.6174, pp. 1006–1010.

Mesoudi, Alex, Andrew Whiten, and Kevin N Laland (2006). "Towards a unified science of cultural evolution". In: *Behavioral and brain sciences* 29.4, pp. 329–347.

Mitteroecker, Philipp and Fred Bookstein (2007). "The conceptual and statistical relationship between modularity and morphological integration". In: *Systematic Biology* 56.5, pp. 818–836.

Mitteroecker, Philipp and Philipp Gunz (2009). "Advances in geometric morphometrics". In: *Evolutionary Biology* 36.2, pp. 235–247.

Mitteroecker, Philipp, Philipp Gunz, and Fred L Bookstein (2005). "Heterochrony and geometric morphometrics: a comparison of cranial growth in *Pan paniscus* versus *Pan troglodytes*". In: *Evolution & development* 7.3, pp. 244–258.

Moisik, Scott Reid and Dan Dediu (2017). "Anatomical biasing and clicks: Evidence from biomechanical modeling". In: *Journal of Language Evolution* 2.1, pp. 37–51.

Mooers, Arne O and Stephen B Heard (1997). "Inferring evolutionary process from phylogenetic tree shape". In: *The quarterly review of Biology* 72.1, pp. 31–54.

Mooshammer, Christine and Christian Geng (2008). "Acoustic and articulatory manifestations of vowel reduction in German". In: *Journal of the International Phonetic Association* 38.2, pp. 117–136.

Moran, Steven and Daniel McCloy, eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.

Moran, Steven, Daniel McCloy, and Richard Wright (2012). "Revisiting Population Size vs. Phoneme Inventory Size". In: *Language* 88.4, pp. 877–893.

Moran, Steven et al. (2024). "Interindividual vocal tract diversity influences the phonetic diversification of spoken languages". In: *bioRxiv*, pp. 2024–07.

Morton, Eugene S (1975). "Ecological sources of selection on avian sounds". In: *The American Naturalist* 109.965, pp. 17–34.

Moseley, Dana Lynn et al. (2018). "Acoustic adaptation to city noise through vocal learning by a songbird". In: *Proceedings of the Royal Society B* 285.1888, p. 20181356.

Müller, Gerd B (2007). "Evo–devo: extending the evolutionary synthesis". In: *Nature reviews genetics* 8.12, pp. 943–949.

Munhall, KEVIN G, Mitsuo Kawato, and Eric Vatikiotis-Bateson (2000). "Coarticulation and physical models of speech production". In: *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, pp. 9–28.

Munson, Benjamin and Molly Babel (2019). "The phonetics of sex and gender". In: *The Routledge Handbook of Phonetics*. Routledge, pp. 499–525.

Namasivayam, Aravind Kumar et al. (2020). "Speech sound disorders in children: An articulatory phonology perspective". In: *Frontiers in Psychology* 10, p. 2998.

Narayanan, Shrikanth et al. (2014). "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)". In: *The Journal of the Acoustical Society of America* 136.3, pp. 1307–1311.

Neiberg, Daniel, Gopal Ananthakrishnan, and Olov Engwall (2008). "The acoustic to articulation mapping: Non-linear or non-unique?" In: *INTERSPEECH*. Vol. 2008, 9th.

Nelson, Douglas A and Peter Marler (1994). "Selection-based learning in bird song development." In: *Proceedings of the National Academy of Sciences* 91.22, pp. 10498–10501.

Nenadić, Filip et al. (2020). "Perception of vowels with missing formant peaks". In: *The Journal of the Acoustical Society of America* 148.4, pp. 1911–1921.

Newton, Ian (2003). *Speciation and biogeography of birds*. Academic Press.

Nichols, Johanna (1992). *Linguistic diversity in space and time*. University of Chicago Press.

Niedzielski, Nancy (1999). "The Effect of Social Information on the Perception of Sociolinguistic Variables". In: *Journal of Language and Social Psychology* 18.1, pp. 62–85.

Nishimura, Takeshi et al. (2022). "Evolutionary loss of complexity in human vocal anatomy as an adaptation for speech". In: *Science* 377.6607, pp. 760–763.

Noiray, Aude, Khalil Iskarous, and Douglas H Whalen (2014). "Variability in English vowels is comparable in articulation and acoustics". In: *Laboratory phonology* 5.2, pp. 271–288.

Nolan, Francis and Tomasina Oh (1996). "Identical twins, different voices". In: *The International Journal of Speech, Language and the Law* 3.1, pp. 39–49.

Nottebohm, Fernando (1969). "The song of the chingolo, *Zonotrichia capensis*, in Argentina: Description and evaluation of a system of dialects". In: *The Condor* 71.3, pp. 299–315.

Novembre, John et al. (2008). "Genes mirror geography within Europe". In: *Nature* 456.7218, pp. 98–101.

Oganian, Yulia et al. (2022). "Vowel and formant representation in human auditory speech cortex". In: *Neuron* 111, pp. 1–14.

Ohala, John J (1989). "Sound change is drawn from a pool of synchronic variation". In: *Language change: Contributions to the study of its causes*, pp. 173–198.

— (1993). "Coarticulation and phonology". In: *Language and speech* 36.2-3, pp. 155–170.

Öhman, Sven EG (1966). "Coarticulation in VCV utterances: Spectrographic measurements". In: *The Journal of the Acoustical Society of America* 39.1, pp. 151–168.

— (1967). "Numerical model of coarticulation". In: *The Journal of the Acoustical Society of America* 41.2, pp. 310–320.

Olsen, Aaron M and Mark W Westneat (2015). "StereoMorph: An R package for the collection of 3D landmarks and curves using a stereo camera set-up". In: *Methods in Ecology and Evolution* 6.3, pp. 351–356.

Pagel, Mark (2017). "Darwinian perspectives on the evolution of human languages". In: *Psychonomic bulletin & review* 24, pp. 151–157.

Pakendorf, Brigitte (2014). "Coevolution of languages and genes". In: *Current opinion in genetics & development* 29, pp. 39–44.

Papcun, George et al. (1992). "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data". In: *The Journal of the Acoustical Society of America* 92.2, pp. 688–700.

Parravicini, Andrea and Telmo Pievani (2018). "Continuity and discontinuity in human language evolution: putting an old-fashioned debate in its historical perspective". In: *Topoi* 37.2, pp. 279–287.

Pastore, Aldo et al. (2022). "Speech listening entails neural encoding of invisible articulatory features". In: *Neuroimage* 264, p. 119724.

Pawlowitsch, Christina, Panayotis Mertikopoulos, and Nikolaus Ritt (2011). "Neutral stability, drift, and the diversification of languages". In: *Journal of theoretical biology* 287, pp. 1–12.

Payne, Robert B (1978). *Population structure and social behavior: models for testing the ecological significance of song dialects in birds*. Museum of Zoology, University of Michigan.

Pericliev, Vladimir (2004). "There Is No Correlation Between the Size of a Community Speaking a Language and the Size of the Phonological Inventory of That Language". In: *Linguistic Typology* 8.3, pp. 376–383.

Perkell, Joseph S and Dennis H Klatt (2014). *Invariance and variability in speech processes*. Psychology Press.

Perrier, Pascal and Susanne Fuchs (2015). "Motor equivalence in speech production". In: *The handbook of speech production*, pp. 223–247.

Pfenning, Andreas R et al. (2014). "Convergent transcriptional specializations in the brains of humans and song-learning birds". In: *Science* 346.6215, p. 1256846.

Pinker, Steven and Paul Bloom (1990). "Natural language and natural selection". In: *Behavioral and brain sciences* 13.4, pp. 707–727.

Pisanski, Katarzyna et al. (2014). "Vocal indicators of body size in men and women: a meta-analysis". In: *Animal Behaviour* 95, pp. 89–99.

Plomp, Reinier, Louis CW Pols, and John P Van de Geer (1967). "Dimensional analysis of vowel spectra". In: *The Journal of the Acoustical Society of America* 41.3, pp. 707–712.

Podos, Jeffrey and Paige S Warren (2007). "The evolution of geographic variation in birdsong". In: *Advances in the Study of Behavior* 37, pp. 403–458.

Ponce de León, Marcia S et al. (2018). "Human bony labyrinth is an indicator of population history and dispersal from Africa". In: *Proceedings of the National Academy of Sciences* 115.16, pp. 4128–4133.

Ponce de León, Marcia S et al. (2021). "The primitive brain of early Homo". In: *Science* 372.6538, pp. 165–171.

Popat, Hashmat et al. (2013). "A geometric morphometric approach to the analysis of lip shape during speech: development of a clinical outcome measure". In: *PLOS one* 8.2, e57368.

Potvin, Dominique A and Sonya M Clegg (2015). "The relative roles of cultural drift and acoustic adaptation in shaping syllable repertoires of island bird populations change with time since colonization". In: *Evolution* 69.2, pp. 368–380.

Preisig, Basil C, Lars Riecke, and Alexis Hervais-Adelman (2022). "Speech sound categorization: The contribution of non-auditory and auditory cortical regions". In: *NeuroImage* 258, p. 119375.

Przybyla, Beata D, Yoshiyuki Horii, and Michael H Crawford (1992). "Vocal fundamental frequency in a twin sample: Looking for a genetic effect". In: *Journal of Voice* 6.3, pp. 261–266.

Purnell, Thomas, William Idsardi, and John Baugh (1999). "Perceptual and phonetic experiments on American English dialect identification". In: *Journal of Language and Social Psychology* 18.1, pp. 10–30.

Puts, David Andrew, Steven JC Gaulin, and Katherine Verdolini (2006). "Dominance and the evolution of sexual dimorphism in human voice pitch". In: *Evolution and Human Behavior* 27.4, pp. 283–296.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rama, Taraka et al. (2018). "Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?" In: *arXiv preprint arXiv:1804.05416*.

Raup, David M (1966). "Geometric analysis of shell coiling: general problems". In: *Journal of paleontology*, pp. 1178–1190.

Rebernik, Teja et al. (2021). "A review of data collection practices using electromagnetic articulography". In: *Laboratory Phonology* 12.1, p. 6.

Recuerda, María et al. (2021). "Sequential colonization of oceanic archipelagos led to a species-level radiation in the common chaffinch complex (Aves: *Fringilla coelebs*)". In: *Molecular Phylogenetics and Evolution* 164, p. 107291.

Relethford, John (2004). "Global patterns of isolation by distance based on genetic and morphological data". In: *Human biology* 76.4, pp. 499–513.

Renfrew, Colin, April McMahon, Robert Lawrence Trask, et al. (2000). "Time depth in historical linguistics". In: *(No Title)*.

Riebel, Katharina, Robert F Lachlan, and Peter JB Slater (2015). "Learning and cultural transmission in chaffinch song". In: *Advances in the Study of Behavior*. Vol. 47. Elsevier, pp. 181–227.

Rivera, Moises et al. (2023). "Machine learning and statistical classification of birdsong link vocal acoustic features with phylogeny". In: *Scientific Reports* 13.1, p. 7076.

Rohlf, F James and Marco Corti (2000). "Use of two-block partial least-squares to study covariation in shape". In: *Systematic Biology* 49.4, pp. 740–753.

Rohlf, F James and Dennis Slice (1990). "Extensions of the Procrustes method for the optimal superimposition of landmarks". In: *Systematic Biology* 39.1, pp. 40–59.

Rojas, Sandra, Elaina Kefalianos, and Adam Vogel (2020). "How does our voice change as we age? A systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 years of age". In: *Journal of Speech, Language, and Hearing Research* 63.2, pp. 533–551.

Rothstein, Stephen I and Robert C Fleischer (1987). "Vocal dialects and their possible relation to honest status signalling in the brown-headed cowbird". In: *The Condor* 89.1, pp. 1–23.

Rousset, Isabelle (2003). "From lexical to syllabic organization: Favored and disfavored co-occurrences". In: *15th International Congress of Phonetic Sciences (ICPhS-15)*.

Saltzman, Elliot L. and Kevin G. Munhall (1989). "A dynamical approach to gestural patterning in speech production". In: *Ecological Psychology* 1.4, pp. 333–382.

Salwiczek, Lucie H and Wolfgang Wickler (2004). "Birdsong: An evolutionary parallel to human language". In: *Semiotica*.

Sandler, Wendy and Diane Carolyn Lillo-Martin (2006). *Sign language and linguistic universals*. Cambridge University Press.

Scharinger, Mathias, William J Idsardi, and Samantha Poe (2011). "A comprehensive three-dimensional cortical map of vowel space". In: *Journal of Cognitive Neuroscience* 23.12, pp. 3972–3982.

Schlager, Stefan (2017). "Morpho and Rvcg – Shape Analysis in R". In: *Statistical Shape and Deformation Analysis*. Ed. by Guoyan Zheng, Shuo Li, and Gabor Szekely. Academic Press, pp. 217–256.

Schliep, Klaus (2011). "phangorn: phylogenetic analysis in R". In: *Bioinformatics* 27.4, pp. 592–593.

Schmandt-Besserat, Denise (2014). "The evolution of writing". In: *International encyclopedia of social and behavioral sciences*, pp. 1–15.

Schneider, Caroline A, Wayne S Rasband, and Kevin W Eliceiri (2012). "NIH Image to ImageJ: 25 years of image analysis". In: *Nature Methods* 9.7, pp. 671–675.

Schwartz, Jean-Luc et al. (1997). "The dispersion-focalization theory of vowel systems". In: *Journal of phonetics* 25.3, pp. 255–286.

Seddon, Nathalie (2005). "Ecological adaptation and species recognition drives vocal evolution in neotropical suboscine birds". In: *Evolution* 59.1, pp. 200–215.

Serrurier, Antoine and Pierre Badin (2005). "Towards a 3D articulatory model of velum based on MRI and CT images". In: *ZAS Papers in Linguistics (Speech production and perception: Experimental analyses and models)* 40, pp. 195–211.

Serrurier, Antoine and Christiane Neuschaefer-Rube (2023). "Morphological and acoustic modeling of the vocal tract". In: *The Journal of the Acoustical Society of America* 153.3, pp. 1867–1886.

Simonyan, Kristina and Barry Horwitz (2011). "Laryngeal motor cortex and control of speech in humans". In: *The Neuroscientist* 17.2, pp. 197–208.

Simpson, Adrian P (2002). "Gender-specific articulatory–acoustic relations in vowel sequences". In: *Journal of Phonetics* 30.3, pp. 417–435.

— (2009). "Phonetic differences between male and female speech". In: *Language and Linguistics Compass* 3.2, pp. 621–640.

Slater, Peter JB, Frederick A Clements, and DJ Goodfellow (1984). "Local and regional variations in chaffinch song and the question of dialects". In: *Behaviour* 88.1-2, pp. 76–97.

Slatkin, Montgomery (1993). "Isolation by distance in equilibrium and non-equilibrium populations". In: *Evolution* 47.1, pp. 264–279.

Slatkin, Montgomery and Laurent Excoffier (2012). "Serial founder effects during range expansion: a spatial analog of genetic drift". In: *Genetics* 191.1, pp. 171–181.

Sorensen, Tanner et al. (2017). "Database of Volumetric and Real-Time Vocal Tract MRI for Speech Science." In: *Interspeech*, pp. 645–649.

Starostin, Sergei (2000). "Comparative-historical linguistics and lexicostatistics". In: *Time depth in historical linguistics* 1, pp. 223–265.

Stephens, SS and J Volkman (1940). "The relation of pitch to frequency". In: *American Journal of Psychology* 53.3, pp. 329–353.

Stevens, Kenneth N (1972). "The quantal nature of speech: Evidence from articulatory-acoustic data". In: *Human Communication: A Unified View*. Ed. by Edward E David and Peter B Denes. McGraw-Hill.

— (2000). *Acoustic phonetics*. Vol. 30. MIT press.

Stevens, Stanley Smith, John Volkmann, and Edwin Broomell Newman (1937). "A scale for the measurement of the psychological magnitude pitch". In: *The journal of the acoustical society of america* 8.3, pp. 185–190.

Stone, Maureen (1997). "Laboratory techniques for investigating speech articulation". In: *The handbook of phonetic sciences* 1, pp. 1–32.

Strauss, Richard E and Fred L Bookstein (1982). "The truss: body form reconstructions in morphometrics". In: *Systematic Biology* 31.2, pp. 113–135.

Suthers, Roderick (Dec. 2004). "How birds sing and why it matters". In: Elsevier, pp. 272–295.

Tienkamp, Thomas B et al. (2024). "The Impact of Electromagnetic Articulography Sensors on the Articulatory-Acoustic Vowel Space in Speakers with and without

Parkinson's Disease". In: *Proceedings of the 13th International Seminar on Speech Production*.

Tilsen, Sam (2013). "Inhibitory mechanisms in speech planning maintain and maximize contrast". In: *Origins of sound change: Approaches to phonologization*, pp. 112–127.

Tiwari, Vibha (2010). "MFCC and its applications in speaker recognition". In: *International journal on emerging technologies* 1.1, pp. 19–22.

Tracey, Brian et al. (2023). "Towards interpretable speech biomarkers: exploring MFCCs". In: *Scientific Reports* 13.1, p. 22787.

Trask, Robert Lawrence (2013). *The history of Basque*. Routledge.

Traunmüller, Hartmut (1990). "Analytical expressions for the tonotopic sensory scale". In: *The Journal of the Acoustical Society of America* 88.1, pp. 97–100.

Tuyl, Rory Van and Asya Pereltsvaig (2012). "Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa"". In: *Science* 335.6069, pp. 657–657.

Ulbaek, Ib (1998). "The origin of language and cognition". In: *Approaches to the Evolution of Language*, pp. 30–43.

Urban, Matthias and Steven Moran (2021). "Altitude and the distributional typology of language structure: Ejectives and beyond". In: *PLOS One* 16.2, pp. 1–36.

Uyeda, Josef C, Rosana Zenil-Ferguson, and Matthew W Pennell (2018). "Rethinking phylogenetic comparative methods". In: *Systematic Biology* 67.6, pp. 1091–1109.

Van den Berg, Janwillem (1958). "Myoelastic-aerodynamic theory of voice production". In: *Journal of speech and hearing research* 1.3, pp. 227–244.

Vaux, Bert and Bridget Samuels (2015). "Explaining vowel systems: dispersion theory vs natural selection". In: *The Linguistic Review* 32.3, pp. 573–599.

Vihman, Marilyn May (2014). *Phonological development: The first two years*. Wiley.

Viswanathan, Navin, James S Magnuson, and Carol A Fowler (2010). "Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech." In: *Journal of Experimental Psychology: Human Perception and Performance* 36.4, p. 1005.

Vorperian, Houri K. et al. (2011). "Developmental Sexual Dimorphism of the Oral and Pharyngeal Portions of the Vocal Tract: An Imaging Study". In: *Journal of Speech, Language, and Hearing Research* 54.4, pp. 995–1010.

Warren, Joe D, AR Jennings, and Timothy D Griffiths (2005). "Analysis of the spectral envelope of sounds by the human brain". In: *Neuroimage* 24.4, pp. 1052–1057.

Weckwerth, Jarosław (2021). "Vowels". In: *The Cambridge Handbook of Phonetics*. Ed. by Rachael-Anne Knight and JaneEditors Setter. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 40–64.

Weirich, Melanie and Susanne Fuchs (2011). "Vocal tract morphology can influence speaker specific realisations of phonemic contrasts". In: *Proceedings of the ISSP*, pp. 251–258.

Weirich, Melanie and Adrian P Simpson (2014). "Articulatory vowel spaces of male and female speakers". In: *Proceedings of the 10th International Seminar on Speech Production*. University of Cologne Cologne, Germany, pp. 449–452.

Wichmann, Søren, Eric W Holman, and Cecil H Brown (2022). "The ASJP database (version 20)". In: *Leipzig: Max Planck Institute for Evolutionary Anthropology*.

Wichmann, Søren et al. (2010). "Evaluating linguistic distance measures". In: *Physica A: Statistical Mechanics and its Applications* 389.17, pp. 3632–3639.

Wilson, Ian (2014). "Using ultrasound for teaching and researching articulation". In: *Acoustical Science and Technology* 35.6, pp. 285–289.

Wood, Sidney (1982). "X-ray and model studies of vowel articulation". In: *Working papers/Lund University, Department of Linguistics and Phonetics* 23.

Wrench, Alan and Janet Beck (2021). "Physiological Foundations". In: *The Cambridge Handbook of Phonetics*. Ed. by Rachael-Anne Knight and Jane Setter. Cambridge University Press, pp. 11–39.

Xu, Anqi et al. (2021). "Model-based exploration of linking between vowel articulatory space and acoustic space". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 4. ISCA, pp. 3191–3195.

Yang, Byunggon (2021). "Measuring Vowels". In: *The Cambridge Handbook of Phonetics*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 261—284.

Yang, Zetian and Michael A Long (2025). "Convergent vocal representations in parrot and human forebrain motor networks". In: *Nature* 640.8058, pp. 427–434.

Yu, Alan C (2021). "Toward an individual-difference perspective on phonologization". In: *Glossa: a journal of general linguistics* 6.1.

Zahorian, Stephen A and Amir Jalali Jagharghi (1993). "Spectral-shape features versus formants as acoustic correlates for vowels". In: *The Journal of the Acoustical Society of America* 94.4, pp. 1966–1982.

Zelditch, Miriam L, Donald L Swiderski, and HH Sheets (2012). *Geometric Morphometrics for Biologists: A Primer*. Elsevier Academic Press.

Zhang, Menghan and Tao Gong (2022). "Structural variability shows power-law based organization of vowel systems". In: *Frontiers in Psychology* 13, p. 801908.

Zhang, Yutao et al. (2023). "Analogies of human speech and bird song: From vocal learning behavior to its neural basis". In: *Frontiers in Psychology* 14, p. 1100969.

Zollikofer, Christoph P. E. et al. (2011). "MRI-based morphometric analysis of the human vocal tract during speech formation and implications for fossil hominin vocal abilities". In: *American Journal of Physical Anthropology*. Vol. 144, pp. 318–318.

Zollikofer, Christoph P. E. et al. (2022). "Endocranial ontogeny and evolution in early Homo sapiens: The evidence from Herto, Ethiopia". In: *Proceedings of the National Academy of Sciences* 119.32, e2123553119.

Zollikofer, Christoph P.E. and Marcia S Ponce de León (2002). "Visualizing patterns of craniofacial shape variation in Homo sapiens". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269.1493, pp. 801–807.

Zsiga, Elizabeth C (2013). *The Sounds of Language: An Introduction to Phonetics and Phonology*. John Wiley & Sons.

Zwicker, Eberhard (1961). "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)". In: *The Journal of the Acoustical Society of America* 33.2, pp. 248–248.