ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Statistics in Phonetics

Shahin Tavakoli,[1] Beatrice Matteo,[1] Davide Pigoli,[2] Eleanor Chodroff,[3] John Coleman,[4] Michele Gubian,[5] Margaret E.L. Renwick,[6,7] and Morgan Sonderegger[8]

[1]Research Institute for Statistics and Information Sciences, Geneva School of Economics and Management, Université de Genève, Geneva, Switzerland; email: shahin.tavakoli@unige.ch, beatrice.matteo@unige.ch

[2]Department of Mathematics, King's College London, London, United Kingdom; email: davide.pigoli@kcl.ac.uk

[3]Department of Computational Linguistics, University of Zurich, Zurich, Switzerland; email: eleanor.chodroff@uzh.ch

[4]Phonetics Laboratory, University of Oxford, Oxford, United Kingdom; email: john.coleman@phon.ox.ac.uk

[5]Institute for Phonetics and Speech Processing (IPS), Ludwig Maximilian University of Munich, Munich, Germany; email: m.gubian@phonetik.uni-muenchen.de

[6]Department of Linguistics, University of Georgia, Athens, Georgia, USA; email: mrenwick@uga.edu

[7]Department of Cognitive Science, Johns Hopkins University, Baltimore, Maryland, USA

[8]Department of Linguistics, McGill University, Montreal, Quebec, Canada; email: morgan.sonderegger@mcgill.ca

Annu. Rev. Stat. Appl. 2025. 12:133–56

First published as a Review in Advance on October 1, 2024

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-112723-034642

Keywords

acoustic phonetics, articulatory phonetics, functional data analysis, generalized mixed-effects models, sociophonetics, speech perception

Abstract

Phonetics is the scientific field concerned with the study of how speech is produced, heard, and perceived. It abounds with data, such as acoustic speech recordings, neuroimaging data, and articulatory data. In this article, we provide an introduction to different areas of phonetics (acoustic phonetics, sociophonetics, speech perception, articulatory phonetics, speech inversion, sound change, and speech technology), an overview of the statistical methods for analyzing their data, and an introduction to the signal processing methods commonly applied to speech recordings. A major transition in the statistical modeling of phonetic data has been the shift from fixed effects to random effects regression models, the modeling of curve data (for instance, via generalized additive mixed models or functional data analysis methods), and the use of Bayesian methods. This shift has been driven in part by the increased focus on large speech corpora in phonetics, which has arisen from

machine learning methods such as forced alignment. We conclude by identifying opportunities for future research.

1. INTRODUCTION

Speech, a fundamental form of human communication, is a very complex process that typically develops naturally from a young age. Phonetics is the primary scientific discipline that studies speech, including each aspect of speech communication shown in **Figure 1**.

Phonetics can be thought of as a subfield of or closely allied field to linguistics, and it also has close ties to physiology, psychology, physics, machine learning, and sociology, among other domains. As nicely expressed by Ohala (2006, p. 468):

> Phonetics attempts to provide answers to such questions as: What is the physical nature and structure of speech? How is speech produced and perceived? How can one best learn to pronounce the sounds of another language? How do children first learn the sounds of their mother tongue? How can one find the cause and the therapy for defects of speech and hearing? How and why do speech sounds vary—in different styles of speaking, in different phonetic contexts, over time, over geographical regions? How can one design optimal mechanical systems to code, transmit, synthesize, and recognize speech?

The acoustic signals produced in speech reflect the sounds of a language, which exist when the language is spoken or when audio recordings of speech are replayed. Sounds change historically: At a given time, older speakers speak differently from younger generations, and the speech of each successive age group changes through history. Sounds are also personal: They reflect aspects of a speaker's social identity (such as age or gender), geographical dialect, and social group. Speech production requires complex and rapid coordinated movements of various speech organs (such as lungs, vocal folds, tongue, and lips; see **Figure 2**), which the typical person takes for granted until they are made aware of them. The inventory of speech sounds and the articulatory gestures used to produce them are not the same across languages or dialects. Speech production and speech perception are typically harder in a second language that is not learned in early childhood, though this is very much a matter of practice, motivation, and ability.

In this review, we focus on statistical methods used in the analysis of speech data across several areas of phonetics, which we briefly introduce. Given the broad scope of the article, we cannot cover all of phonetics or all statistical methods used in phonetics. Additional coverage is given

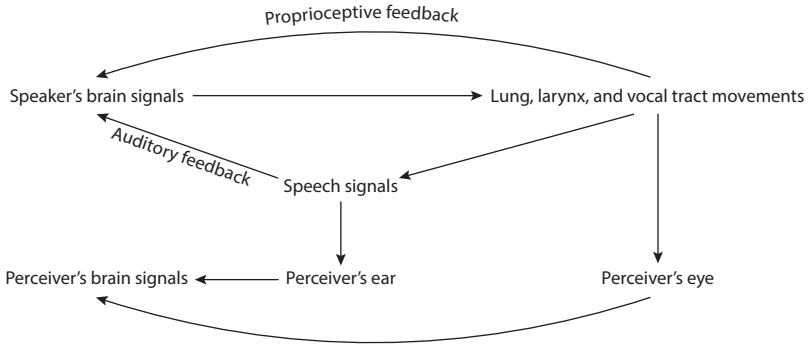

Figure 1

The speech pipeline. Each arrow indicates one aspect of speech communication.

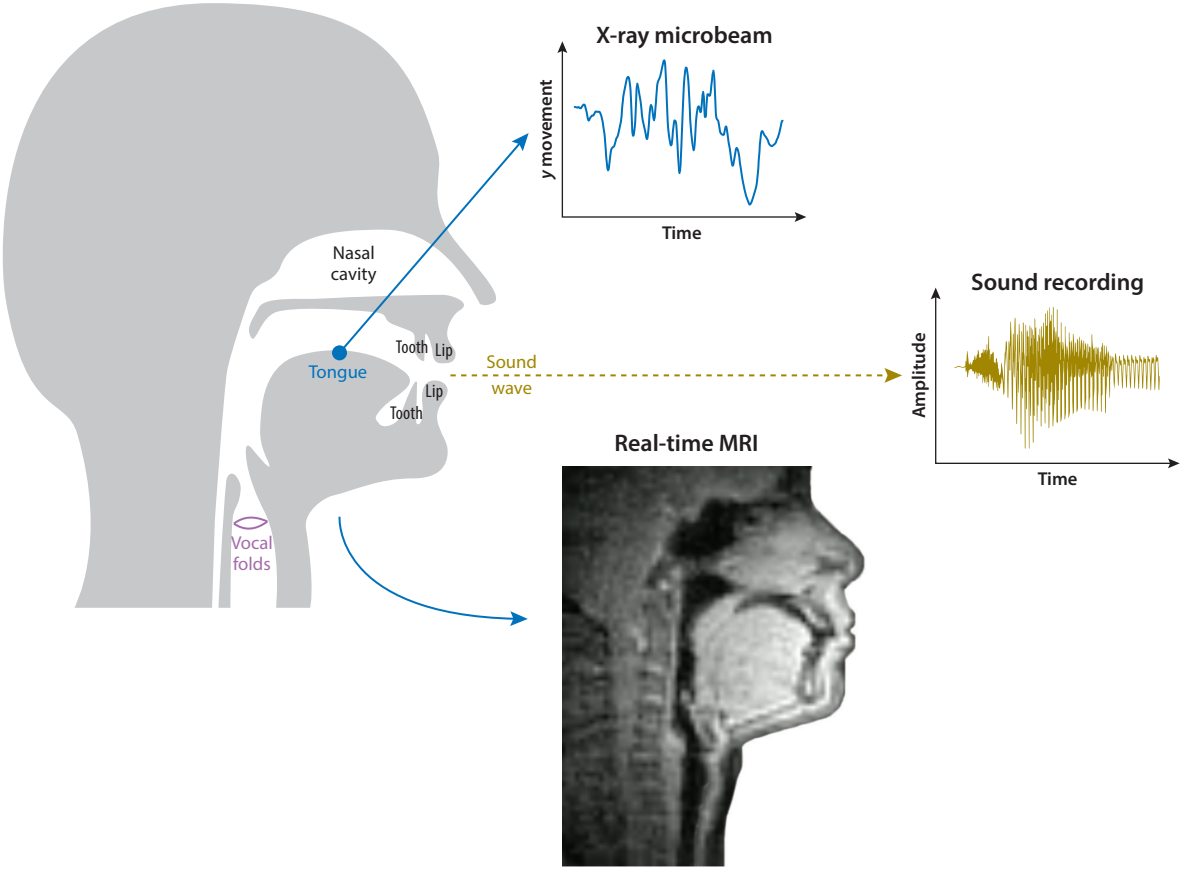

Figure 2

Some organs of speech and examples of data that can be measured from them. The X-ray microbeam data are from Westbury et al. (1994). Real-time magnetic resonance imaging (MRI) data are from Lim et al. (2021). We thank Wugapodes for the unlabeled version of the midsagittal head diagram, shared under CC0 1.0 (**https://commons.wikimedia.org/wiki/File:Midsagittal_diagram_unlabeled.svg**).

by Sonderegger & Sóskuthy (2024) in a review of statistical practice in phonetics written for phoneticians and speech scientists, whose approach is complementary to the current article.

We briefly describe the core concepts of speech production (Section 2), then turn to representations of sound recordings (Section 3). The core of the article is Section 4, where we link different areas of phonetics to the statistical methods used to model specific types of phonetic data. Section 4.1 introduces acoustic phonetics and related topics (speech variation, sociophonetics), followed in Section 4.2 by speech perception. We then turn to articulatory phonetics (Section 4.3) and speech inversion (Section 4.4). Sound change is presented in Section 4.5, and Section 4.6 treats the increasing role of machine learning in phonetics. Section 5 concludes with some open questions.

2. BASICS OF SPEECH PRODUCTION

To produce most speech sounds, humans expel air from the lungs. Air first passes through the vocal folds, which can remain open, be constricted, or close completely. When they remain open, the

resulting sound is called voiceless or unvoiced. When they constrict, the vocal folds may vibrate, and the resulting sound is called voiced. The air, and the sound waves propagating through it, then passes through the vocal tract: first through the pharynx and then through either the oral cavity or the nasal cavity, or both. The tongue modulates the sound as it passes through the oral cavity by making constrictions, where the tongue comes into close proximity or contact with parts of the oral cavity, thereby producing various sounds. The shape of the lips also affects the produced sound. The organs of speech, as well as examples of speech data, are depicted in **Figure 2**.

Hertz (Hz): a unit of frequency equal to one cycle per second

Kilohertz (kHz): a unit of frequency equal to 1,000 Hz

The usual way to model the relationship between articulation, described above, and acoustics, the sound that is output at the lips, is the source/filter model of speech production (Fant 1960, 1980; Stevens 1998). In this framework, voiced speech sounds start with the vocal folds vibrating, and the resulting sound wave passes through the vocal tract, which essentially acts as a filter, amplifying some frequencies and damping others. A maximally simplified model for the filter is the tube model (Fant 1960, Johnson 2012). It models the vocal tract as a series of connected tubes, either open or closed at their ends. From physics, we know that different lengths and diameters of the tubes give rise to different resonant frequencies, called formants, which are commonly used by phoneticians to characterize the produced speech sounds. While the tube model is an idealized model of speech production, it is a surprisingly good first approximation that explains how configurations of the vocal tract generate different speech sounds. Crucially, the tube model also motivates looking at sound recordings through the lens of Fourier analysis (see Section 3).

3. REPRESENTATIONS OF THE SOUND WAVE AND DERIVED FEATURES

Speech generates air pressure variations that are perceived as sound. To record speech we can use a microphone, which has a diaphragm that vibrates under the action of sound waves, and this vibration is converted to an electrical current. This electrical current is then sampled at a given rate (typically 16 kHz for speech, or 44.1 kHz for the full spectrum of audible sound) and quantized, typically to signed 16-bit amplitude measurements.

A (mono)speech recording can be viewed as a time series x_t, where t is the time in seconds after the start of the recording, for instance $t = 0, 1/16{,}000, 2/16{,}000, \ldots$, if the sampling rate is 16 kHz. **Figure 3*a*** shows an example. In the language of time series, a speech recording is not a stationary process (Shumway & Stoffer 2017), but at the scale of about 10 ms (10^{-2} s), most (but not all) segments are fairly stationary (stop consonants, such as [p] as in "**p**at," may exhibit a short audible burst of about 5 ms duration; here, as is customary, we give phonetic transcriptions of speech sounds in square brackets). For voiced sounds, the time series is (locally) well approximated as a sum of finitely many sine waves. For fricatives (e.g., [s] as in "**s**ea"), the time series is closer to a sum of sine waves over a continuum of frequencies. Given the different characteristics of these sounds in the time domain, a frequency-domain approach to the analysis of speech recordings is often more informative. This approach is very important in the study of speech and language, as discussed from Section 4.1 onward. A useful visual representation of speech recordings is a spectrogram, which is a time-frequency representation of the time series x_t. To analyze the signal locally in time, say using an interval of 10 ms, the squared modulus of the fast Fourier transform (FFT) (Cooley & Tukey 1965), or periodogram, is computed for x_t. This gives the energy of each frequency component of x_t, locally in time. The FFT is repeatedly computed by shifting the time window, resulting in a matrix, where each column corresponds to the periodogram of one time window and each row corresponds to the energy of one frequency range over time. This matrix is plotted by mapping the energy values onto a color (or monochrome) scale, possibly

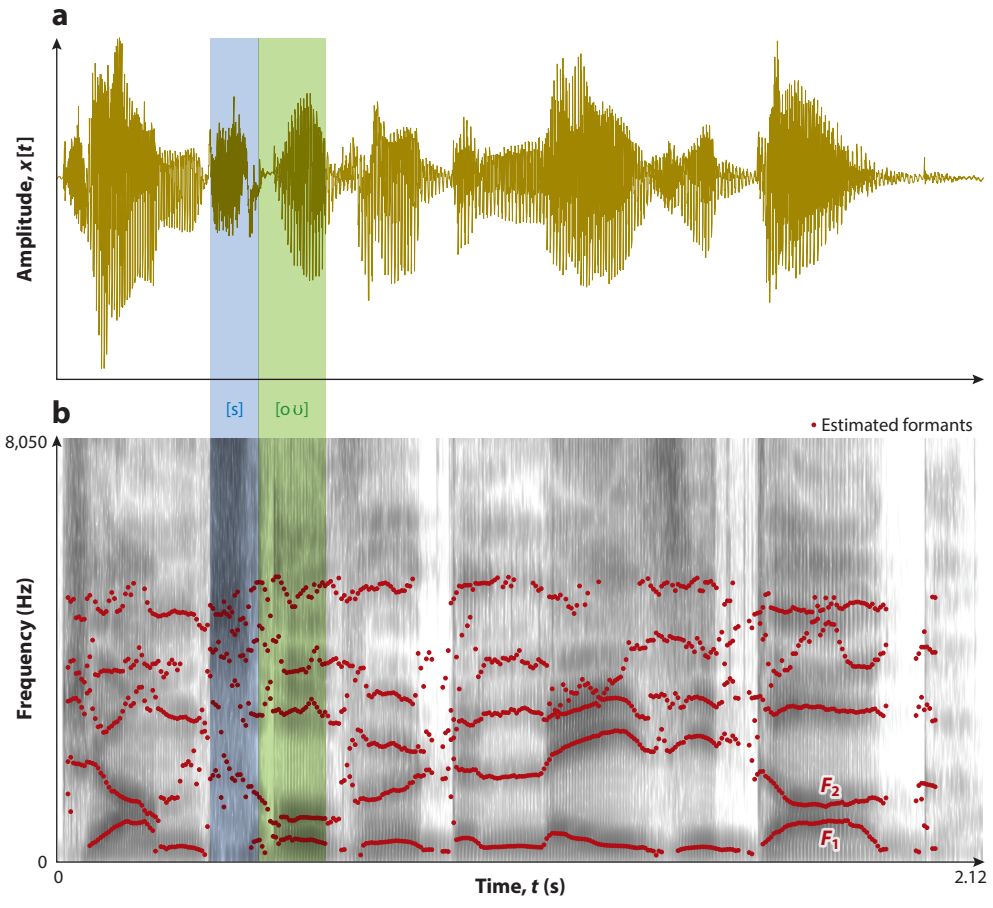

Figure 3

Example of a speech recording x_t of the first author saying "jumps over the lazy dog." (*a*) An oscillogram (a plot of the amplitude x_t against the time t). (*b*) Its wideband spectrogram with estimated formants (*red*). Highlighted in blue is the segment corresponding to the fricative [s] (as in "jumps") and in green the segment corresponding to [oʊ] (as in "over"). The first five formant frequencies are drawn in red on the spectrogram. Some estimated formant frequencies match the spectral peaks in the wideband spectrogram (mostly for the bottom ones, F_1 and F_2), while higher formant frequencies (F_3, F_4, and F_5) sometimes poorly estimate the spectral peaks. The figure was generated using the Praat software (Boersma & Weenink 2023).

after a logarithmic transformation (**Figure 3*b***). Another Fourier transform can be applied to the frequency range of the log-spectrogram, locally in time, resulting in the cepstrum (Bogert et al. 1963). Cepstral analysis is useful as a method for separating the source and the filter (presented in Section 2; see, e.g., Schafer & Rabiner 1970).

The *y* axis of the spectrogram, representing frequencies, may also be transformed nonlinearly into scales that are more closely aligned to human perception, resulting in one of the mel, Bark, or auditory spectrograms (Gold et al. 2011, Johnson 2012). A discrete cosine transform is sometimes applied to the log mel spectrogram (over the frequency bands), resulting in mel-frequency cepstral coefficients (MFCCs). Such coefficients are widely used in automatic speech recognition (ASR) and synthesis (Section 4.6), though they have also been used in studies of speech variation (Section 4.1) and sound change (Section 4.5). **Figure 4** summarizes the computational links between these speech representations.

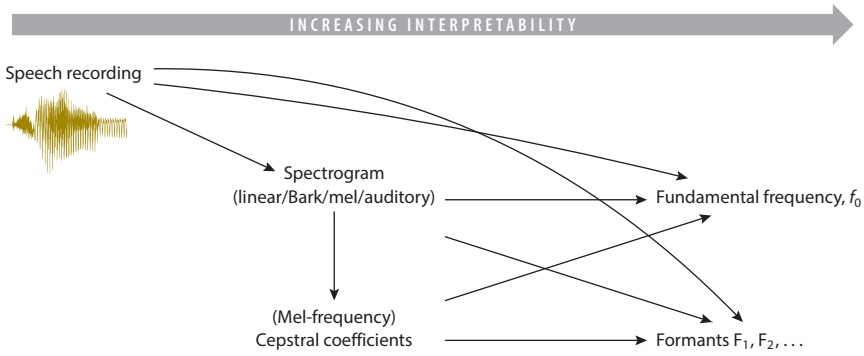

Figure 4

Common speech features/representations and the computational links between them.

Spectral envelope: the spectrum's overall shape, after abstracting away from the small bumps of the harmonics

Wideband spectrogram: a spectrogram obtained by using small time windows, of about 5 ms, giving low frequency resolution but high temporal resolution

Perceptually important acoustic features of voiced speech include the fundamental frequency f_0 and the spectral peaks or formants $F_1, F_2, \ldots$. Each formant has a center frequency ($F_1, F_2, \ldots$), which is particularly important for distinguishing among different kinds of vowels and consonants. Other acoustic features of speech (voiced or unvoiced) that are often analyzed include amplitude and duration, whether of sounds or other units corresponding to intervals of speech, such as syllables, words and phrases. Amplitude is typically measured by the root mean square amplitude (see section 3.3.1 of Johnson 2012); it should not be confused with loudness, which is a perceptual quantity.

The fundamental frequency f_0, not to be confused with pitch (a perceptual property of how we hear and interpret the frequency of a sound), is the distance between the peaks (harmonics) of the periodogram, typically measured in hertz (**Figure 5**). Methods for estimating f_0 include those given by Boersma (1993) and Talkin (1995).

On a given time window, the formant frequencies $F_1 < F_2 < \ldots$ are the frequencies of the peaks of the spectral envelope of the periodogram (see section 4.2 of Coleman 2005), which can be visualized using a wideband spectrogram (**Figure 3*b***). To estimate the formant frequencies,

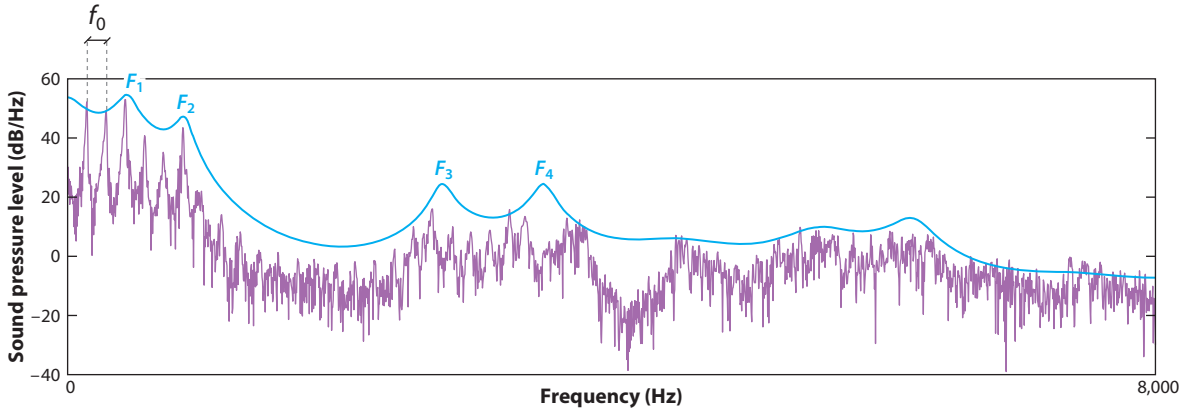

Figure 5

Periodogram (*jagged purple curve*) and LPC spectrum (*smooth blue curve*) of a portion of one speaker's pronunciation of the vowel [oʊ] as in "**o**ver," with annotations for the fundamental frequency f_0 and formant frequencies. The LPC spectrum has been shifted vertically for ease of visualization. Abbreviation: LPC, linear predictive coding.

a technique called linear predictive coding (LPC) (Atal & Schroeder 1978, Schroeder 1985, Coleman 2005) is typically used. In this approach, an autoregressive model of order p, or AR(p) model (see, e.g., section 3.1 of Shumway & Stoffer 2017), is fitted to x_t via the Burg algorithm, which produces stable estimates. The formant frequencies correspond to the peaks of the AR(p) spectral density, also known as the LPC spectrum (**Figure 5**). The number of formants that can be detected is approximately $(p - 2)/2$, depending on the sampling rate used. With a sampling rate of 8 kHz, we can observe frequencies from 0 to 4 kHz, which will show about four formants, so $p = 10$ LPC coefficients is appropriate. With a sampling rate of 16 kHz (8 kHz upper frequency limit), we can observe five or six formants, so $p = 12$ or 14 LPC coefficients is appropriate. Alongside each formant frequency F_j, its bandwidth B_j (a measure of the width of the peak; Stevens 1998, pp. 153–62) is also of interest, because it represents the range of acoustic frequencies excited by the formant and affects the perception of a speaker's voice timbre, among other things.

Voice onset time (VOT): the time interval between the release of a stop consonant and the onset of voicing in the following sound

4. STATISTICAL METHODS

4.1. Acoustic Phonetics

Much research in phonetics falls under the category of acoustic phonetics (Johnson 2012), which studies the acoustic realization of speech sounds, using the representations just discussed, such as spectrograms and formants.

4.1.1. Sociophonetics. In acoustic phonetics, a common problem is to quantify variation in the acoustic realizations of different consonants, vowels, patterns of pitch, or other parameters of interest. For example, researchers may investigate similar vowels across geographical space (i.e., dialects or languages) or between various social groups. The latter is part of sociophonetics, a very active research area that emerged in the twenty-first century (Kendall et al. 2023). Sociophonetics studies how language is used to express social or context-dependent layers of meaning.

Sociophonetic analysis is inspired by sociolinguistic methods, which, beginning in the 1970s, treated spoken language variation in terms of categorical differences in pronunciation, using multivariate analysis programs like GoldVarb and VARBRUL (Cedergren & Sankoff 1974, Tagliamonte 2002). In modern sociophonetics, and in acoustic phonetics more generally, variation is often quantified with continuous phonetic measurements, as described in the following sections.

4.1.2. Analysis of scalar measurements. Many studies focus on the analysis of one or a few scalar measurements. For instance, in the production of stop consonants (e.g., [p]), voice onset time (VOT) (Lisker & Abramson 1964, Cho et al. 2019) is a key feature. Vowel pronunciations are often investigated using the first three formant frequencies F_1, F_2, and F_3 at mid-vowel (Hagiwara 1997). The first two have a nice interpretation in terms of articulation: F_1 is correlated with how much the tongue is lowered, and F_2 is correlated with the location of the tongue–palate constriction (i.e., tongue "backness"), as shown in **Figure 6**.

Normalization of formant frequencies has been much discussed, with the goal of rendering those from different speakers comparable to one another, allowing data to be pooled across people. Normalization is necessary because variation in speakers' physical characteristics (e.g., vocal tract length, mouth size) introduces variability that is usually irrelevant to the identity of the words, vowels, and consonants being spoken and does not disrupt accurate perception by listeners (Johnson & Sjerps 2021). The goal of normalization is to remove such physiologically derived differences while preserving both sociophonetic characteristics (such as dialectal variations, cross-language differences, and group differences) and distinctions between different vowels. In addition to their use as a tool to remove extraneous variation, normalization procedures are often of interest as models of the cognitive processes that allow us to recognize the same vowel when pronounced

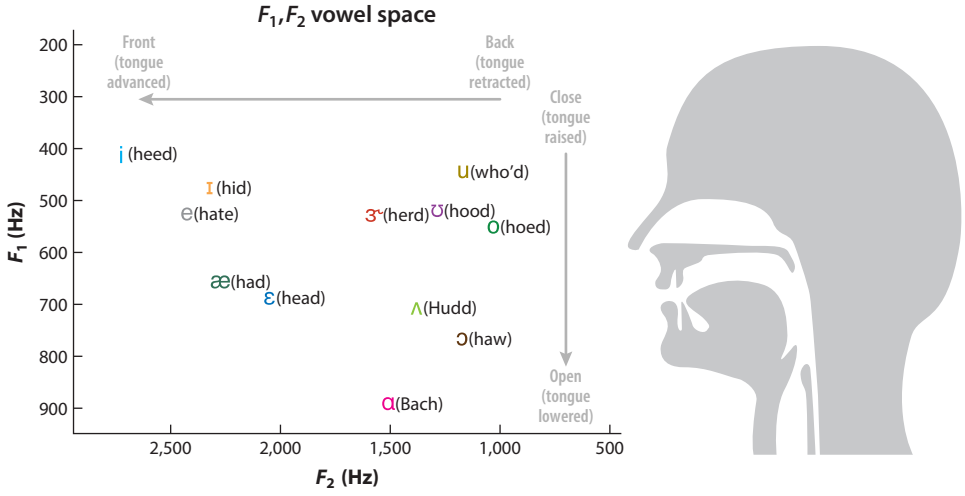

Figure 6

The F_1, F_2 vowel space (*left*) of American English with average (raw) vowel formants over a set of 139 speakers. Each vowel is written in the International Phonetics Alphabet, with an American English pronunciation example in parentheses. The data used are from Hillenbrand et al. (1995) and are available as dataset `h95` in the R package `phonTools` (Barreda 2023). Note that the axis scales increase downward and leftward, in the opposite direction of mathematical convention; this is the tradition for such plots, so that the direction of changes in formant frequencies corresponds to the direction of tongue movements in a left-facing head (*right*). We thank Wugapodes for the diagram on the right, shared under CC0 1.0 (**https://commons.wikimedia.org/wiki/File:Midsagittal_diagram_unlabeled.svg**).

by different speakers (Rosner & Pickering 1994). There are numerous normalization methods, reviewed, for instance, by Voeten et al. (2022). An example is Lobanov's (1971) method, where the formant frequencies are z-scored per speaker—that is, the jth formant frequency for the vowel v is normalized as $F^*_{j,v} = (F_{j,v} - \overline{F}_j)/s_j$, where $\overline{F}_j$ is the average of the formant frequency F_j across all vowels within a speaker and s_j is the standard deviation of the same sample. **Figure 7** shows an example of raw and Lobanov-normalized plots of vowel formant frequencies.

Once normalization has been applied, subsequent statistical analysis aims to explore differences in acoustic parameters relevant to research questions: differences between groups of speakers or words, differences as a function of surrounding sounds, and so on. The statistical methods used initially were often simple visual comparisons, t-tests, and analysis of variance (ANOVA). However, as most phonetic experiments are based on measuring speech characteristics (such as VOT or formants) from a sample of speakers speaking a selection of words, possibly with replications, in the 2000s it was realized that ordinary linear models were often inadequate for these settings. Since the goal of phonetic studies is to make inferences about the general population (either speakers of a certain language or all humans), differences between speakers and words are better modeled with participant and word random effects. Phoneticians from the mid-2000s increasingly used repeated-measures ANOVA, and then around 2010, linear mixed-effects models started to become the norm, due in part to the availability of the R package `lme4` (Bates et al. 2015) and influential books and articles using R code (e.g., Baayen 2008, Baayen et al. 2008); a detailed account is provided by Sonderegger & Sóskuthy (2024). In current practice, (generalized) linear mixed-effects modeling is ubiquitous in phonetic data analysis, and it is the focus of current textbooks covering statistical analysis of phonetic data (e.g., Winter 2019, Sonderegger 2023).

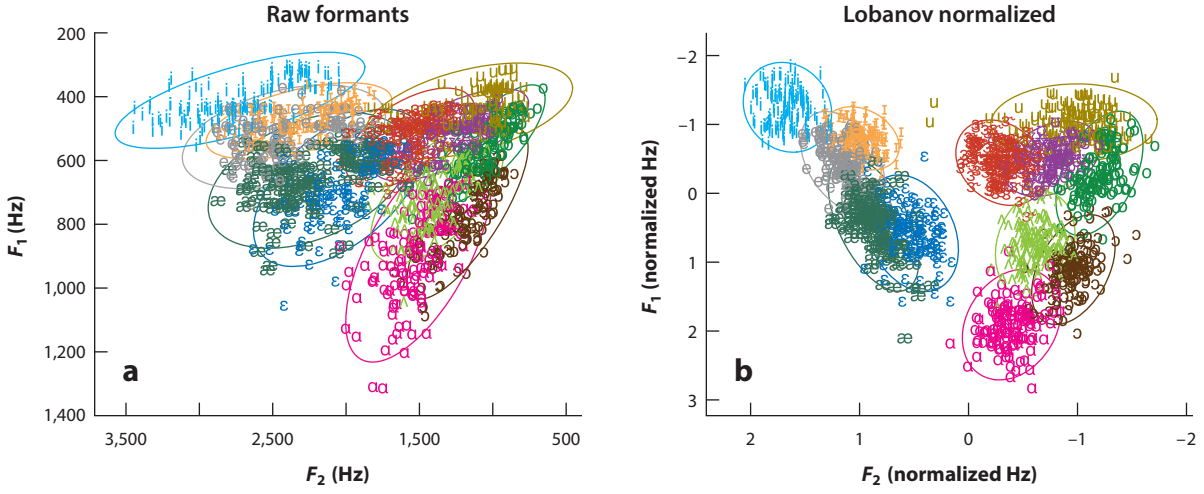

Figure 7

(*a*) Raw vowel space (F_1, F_2) and (*b*) Lobanov-normalized vowel space for data from Hillenbrand et al. (1995) available as dataset `h95` in the R package `phonTools` (Barreda 2023). The superimposed ellipses are approximate 95% confidence regions for each vowel. Note that the axes are reversed, as in **Figure 6**.

The output of (generalized) linear mixed-effects modeling can be used as input to subsequent analyses. For instance, the speaker random intercepts can be used to assess the relative position of a person's speech compared with other speakers of the same age (to detect, for instance, leaders or laggers in sound change), and taken jointly over F_1, F_2 for various vowels, they can be analyzed via principal component analysis to detect systematic covariation between vowels (Brand et al. 2021).

Bayesian methods and philosophy (Gelman et al. 2013) have become more popular in recent years, facilitated by the R package `brms` (Bürkner 2017), a user-friendly interface to the probabilistic programming language Stan (Carpenter et al. 2016). Bayesian models have allowed greater flexibility and the integration of prior information about relevant variables to model speech data (e.g., Tanner et al. 2020, Chodroff & Wilson 2022), in contrast to previous methods that were essentially frequentist (Vasishth et al. 2018). Bayesian models also have the convenient property of providing estimates of random effects when frequentist methods (such as those in `lme4`) fail to converge. However, to the best of our knowledge, it is unknown whether Bayesian estimation is reliable in such situations (though see Eager & Roy 2017). Sonderegger & Sóskuthy (2024, section 3) provide a detailed review of Bayesian methods in phonetics.

4.1.3. Analysis of trajectories. For particular speech sounds, such as diphthongs (e.g., the vowel sound in "loud"), the formant frequencies at mid-vowel alone do not capture the sound's dynamic nature, so it is necessary to model the entire time course of parameters of interest. This led to the use and development of generalized additive mixed models (GAMMs; e.g., Wood 2017, Wieling 2018) and functional data analysis (FDA; Ramsay & Silverman 2005) approaches (Aston et al. 2010, Tavakoli et al. 2019, Renwick & Stanley 2020, Koshy & Tavakoli 2022).

GAMMs can analyze dynamic phonetic data, such as formant frequency trajectories, articulatory movements, or f_0 trajectories, without simplifying or aggregating the data over time, and can flexibly capture the shape of the response curve while accounting for random effects of speakers and items (e.g., vowels). They can also be used to model a nonlinear effect of a covariate on the response, such as the effect of age on formant frequencies (e.g., Fruehwald 2017) and can be fitted in R using the `mgcv` package (Wood 2011). An example of an application of a GAMM with

nonparametric (fixed) effects and additive random intercepts and slopes for speaker, word uttered, and dataset is given by Renwick et al. (2023), modeling change over generational time in the pronunciation of vowels. The `mgcv` implementation allows random variation in curve shape at the speaker and item levels, owing to the link between penalization parameters and random effects (Wood 2017), and a correlation structure in the errors can be included. For an application to the comparison of tongue tip movement between native and nonnative English speakers, readers are directed to Wieling (2018). These models have been used by Chuang et al. (2021) to analyze response time in studies of high-level tones and the ongoing merger[1] of two sets of sibilants in Taiwan Mandarin. However, their complexity poses challenges in terms of inference and interpretation (Thul et al. 2021). Additionally, the computational cost can become high if many basis functions must be used (especially with random variation in the curve shape).

Similar models defined in the framework of functional mixed-effects models (Guo 2002) offer a larger variety of estimation approaches. For example, functional principal components can be used to reduce the number of basis functions in a data-driven way, so that standard estimation procedures for linear mixed-effects models can be used. This approach has been used to model fundamental frequency (f_0) in the tone language Qiang (Aston et al. 2010, Evans et al. 2010) and is available through the R package `multifamm` (Volkmann et al. 2023), which allows the user to jointly model multiple response curves, such as the f_0 and formant frequency trajectories. Other models include functional mixed-effects models for irregularly sampled functional data (Pouplier et al. 2017, with the R package `sparseFLMM`).

More generally, FDA can help capture the time-dependent nature of phonetic features. For example, time-warping (see, e.g., Marron et al. 2015, which discusses in particular the square-root velocity function and the Fisher–Rao metric) can address the problem of time normalization for voice signals (Lucero & Koenig 2000) such that time variability can become part of the statistical analysis. Koenig et al. (2008) explored differences in time and curve-shape variability for oral airflow signals in fricative production between children and adults. Time-warping can also be included in the response of functional mixed-effects models to treat individual variation in signal phase and amplitude, such as in the analysis by Hadjipantelis et al. (2015) of f_0 trajectories in Mandarin Chinese. Other examples include modeling f_0 trajectories with orthogonal polynomials (Grabe et al. 2007), functional principal components for dimension reduction to jointly explore formants and speech rate (Gubian et al. 2015), function-on-scalar regression for analyzing spectra of fricatives and stop releases (Puggaard-Rode 2022), nonparametric functional regression for spatial smoothing (Tavakoli et al. 2019), and generalized linear models with functional predictors to capture differences in pronunciation (Koshy & Tavakoli 2022).

For a more in-depth review of methods for analyzing dynamic data in phonetics (e.g., formant trajectories), readers are directed to Sonderegger & Sóskuthy (2024, section 4).

4.2. Speech Perception

In perception studies, experiments typically measure a subject's response to acoustic speech stimuli (Delattre et al. 1955, Lisker & Abramson 1970, Pisoni & Tash 1974, Kohler 1987). These stimuli are often generated or manipulated artificially, for instance by varying VOT, formant frequencies, or pitch. Responses may be discrete motor responses (e.g., key presses representing listener

[1]Merger occurs when two sounds that were historically pronounced differently in a language come to be pronounced identically. For example, the spellings ⟨ea⟩ and ⟨ee⟩ in English word pairs like meat/meet and feat/feet reflect two distinct vowels in Middle English that are merged in modern English. In Southern US English dialects, the vowels [ɛ] and [ɪ] are merged before nasal consonants, so that words like "pin" and "pen" are pronounced identically.

judgements) or continuous signals (e.g., physiological measurements). The data from such experiments were initially modeled through linear models, followed by linear mixed models, similarly to the analysis of univariate pitch or formant frequency measurements (see Section 4.1). Even though early software for sound manipulation generated odd-sounding or unrealistic sounds, methods that now exist for generating good-quality synthetic speech for use in perception experiments include Klatt synthesis (Klatt & Klatt 1990), LPC-based manipulation of pitch (e.g., Kohler 1987), Tandem-STRAIGHT (Kawahara et al. 2008), and MFCC synthesis (Erro et al. 2014, Hudson et al. 2024).

Other perceptual methodologies expose participants to auditory stimuli while brain activity is measured using, for instance, electroencephalography (EEG) (Ombao et al. 2016) or magnetoencephalography (Proudfoot et al. 2014). The variable of interest is the brain signal change when an atypical auditory stimulus is presented. For each stimulus, event-related potentials (ERPs) (e.g., Kaan 2007) are recorded; these are often averaged at the subject level, and the ERPs of typical stimuli are compared with those of atypical stimuli, often using the mismatch negativity (e.g., Näätänen et al. 1978). Analyzing EEG data involves a standard neuroimaging statistical toolbox (Ombao et al. 2016). For univariate summaries of the EEG signal, the methods used are similar to those described in Section 4.1, with an emphasis on the use of random effects for stimulus and subject effects. Modeling the full multivariate EEG signal (the time series) requires a high level of statistical sophistication, similar to that used in functional magnetic resonance imaging (fMRI) studies (Ombao et al. 2016), in particular in the use of random effects, multiple testing corrections for mass univariate approaches (Groppe et al. 2011), and cluster detection (Frossard & Renaud 2022).

Complementing brain activity measurement, experiments measuring the movements of the eyes—using the visual world paradigm (Barr 2008)—open a window into speech perception and processing as it happens. These experiments involve tracking eye movements of the subject following auditory stimuli and define time series, which are often summarized by the proportion of gaze fixations to areas of interest (such as visual depictions of words beginning with a certain sound). These are then analyzed by a linear mixed-effects model or a repeated-measures ANOVA and growth curve analysis (Mirman et al. 2008). It is an ongoing debate among psycholinguists and phoneticians whether such analyses are sufficiently informative or whether they should analyze the entire trajectories of eye movement (which in principle carry much more information, but are much harder to model), using, for instance, GAMMs (Nixon et al. 2016).

Event-related potential (ERP): the EEG signal following a stimulus

Mismatch negativity: the negative ERP peak in the range 100–250 ms after a stimulus

Visual world paradigm: a type of experiment in which participants hear an auditory stimulus while looking at visual displays of words and distractors

4.3. Articulatory Phonetics

Articulatory phonetics, which studies how speech is produced through the movements of parts of the human vocal tract (see Section 2), plays a key role in understanding the variability in speech production. Many studies focus on the relationship between specific movements of the articulators (**Figure 2**)—such as tongue shapes and lip rounding (e.g., Harshman et al. 1977)—and the acoustic output as seen in a spectrogram (**Figure 3*b***). For example, as illustrated in **Figure 6**, the vertical and horizontal positions of the tongue body roughly determine the first and second formant frequencies. Long-standing questions are how accurate this first approximation is and how other aspects of articulation (like lip rounding) affect formant frequencies.

The available measurement technologies for tracking articulator movements constrain how they can be modeled. Point-tracking methods, such as X-ray microbeam[2] (Kiritani et al. 1975,

[2]X-ray microbeam is a technology in which gold pellets are attached to external and internal articulators on the midsagittal plane, and the movement of the pellets is automatically tracked using movable X-ray beams.

Stone 1990) and electromagnetic articulography (EMA)[3] (Pouplier et al. 2017, Wieling 2018), allow the measurement of particular points on the articulators over time. The advantage of such methods is their high temporal resolution, but the spatial resolution is limited to about six points for EMA; interesting gestures could be missed if they do not appear at those points. In contrast, medical imaging methods provide measurements over a spatial continuum. Ultrasound, especially, is used to study tongue (Davidson 2006, Mielke et al. 2017, Pini et al. 2019) and larynx (Moisik et al. 2014) gestures, but only soft tissue can be reconstructed with this imaging modality, which can be problematic for studying certain speech sounds, such as fricatives. Real-time magnetic resonance imaging (rtMRI) (Toutios & Narayanan 2016) allows the measurement of mid-sagittal-plane images of the vocal tract soft tissues in real time, but at a reasonably high temporal resolution (e.g., a dataset with a temporal resolution of 83 frames per second and a spatial resolution of 84×84 pixels is provided in Lim et al. 2021).

As instrumentation for articulatory measurement is often quite invasive, foundational studies about various aspects of articulation tend to have very few speakers and therefore not much opportunity for comparative statistics, though statistical tests used early on include Student's *t*-test and ANOVA. Later on, phoneticians became more aware of some of the problems of inappropriate ANOVA (e.g., when used without checking for homogeneity of variance, normal errors, and independence of observations), and more sophisticated methods such as repeated-measures ANOVA and MANOVA (multivariate analysis of variance) were used.

The many recent statistical methods used in articulatory phonetics include FDA (Ramsay & Silverman 2005). Indeed, even if the articulatory data are based on a point-tracking method, the time-series nature of the measurements lends itself naturally to viewing the data as multivariate curves. In the case of rtMRI and ultrasound data, the data are of a functional nature in space even for a fixed time point. FDA was introduced to articulatory phonetics by Ramsay et al. (1996), who used a functional principal component analysis and a functional ANOVA to analyze lip motion. Given the subject and item levels present in most articulatory phonetic datasets, random effects need to be included in the model, and they can be included through GAMMs (Tomaschek et al. 2013, Wieling et al. 2016) or FDA approaches, such as by Cederbaum et al. (2016) (see also Carignan et al. 2020). Our view is that these two approaches are very similar, particularly as the R package `mgcv` to fit GAMMs allows the inclusion of smooth random effects in the model (see `factor.smooth` in the R package `mgcv`; Wood 2017, pp. 326–27). The difference lies in the basis functions used to model the curves, as `mgcv` mainly uses spline or similar bases, whereas functional data approaches allow the use of any function basis.

4.4. Speech Inversion

Speech inversion is the estimation of the time series y_t of articulator shapes given a speech recording x_t, that is, the acoustic-to-articulatory inverse map based on inversion of the articulatory-to-acoustic forward map of speech production, depicted in **Figure 8**.

Speech inversion has a variety of applications—in clinical practice, phonetics, language learning, and computer graphics—but it is challenging because it is a nonlinear, ill-posed inverse problem (Kaipio & Somersalo 2005). The acoustic-to-articulatory map is nonlinear and one-to-many; different vocal tract shapes may produce almost identical acoustic output (Atal et al. 1978, Mrayati et al. 1988, Toda et al. 2004). Even in a lossless tube model of speech production, the

[3] EMA employs an instrument, the articulograph, in which coils of wire (which act as sensors) are placed on internal and external parts of the mouth and a variable electromagnetic field is created around the head of the patient, which can be used to estimate the sensors' positions (Rebernik et al. 2021).

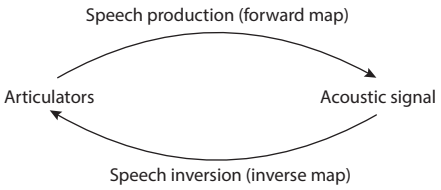

Figure 8

The speech production forward map and speech inversion inverse map.

vocal tract area function is not uniquely determined by the formant frequencies (Mermelstein 1967, Schroeder 1967). The one-to-many aspect of speech inversion comes mainly from the fact that different articulatory constrictions can create similar spectral filters (Mrayati et al. 1988). This happens in part because some constrictions produce antiresonances, which dampen the energy at some frequencies (for instance, in nasal speech sounds; see Johnson 2012, chapter 9). Broadly speaking, the speech inversion problem can be tackled either through acoustic models of the vocal tract or through measurements of real (human-produced) articulatory/acoustic pairs.

Acoustic models of the vocal tract can describe the forward map, in which the articulatory shape is captured by a cross-sectional area of the vocal tract (e.g., Beautemps et al. 1995). This can be used to create an artificial training set of articulatory/acoustic pairs—an articulatory "codebook"—to estimate the inverse mapping (Atal et al. 1978, Larar et al. 1988, Schroeter & Sondhi 1989, Schroeter et al. 1990, Rahim et al. 1991). Alternatively, the forward mapping can be directly used in a hidden Markov model (HMM) (e.g., Murphy 2023, chapter 9), with the hidden chain being the sequence of discrete or continuous articulatory parameters and the observed chain being the acoustic signal (Hiroya & Honda 2004). However, acoustic models of the vocal tract are far from an optimal approximation of the true forward mapping, because the vocal tract is irregular and composed of soft and hard tissue. This leads to energy losses due to yielding vocal tract walls, heat conduction, and viscosity, along with radiation impedances when the sound wave leaves the lips and nose (Fant 1972, Stevens 1998).

Joint measurement of human-produced articulation shapes and associated acoustic signals, as described in Section 4.3, can be used as a training set for estimating the inverse mapping. This is not an easy task: rtMRI, for example, is a very loud procedure, making good audio recordings of speech difficult to obtain. However, the scanner noise may be mitigated using a combination of noise-canceling microphones and postprocessing of the recorded signals (Malinen & Palo 2009).

Nevertheless, Gaussian mixture models (Toda et al. 2004) and nonparametric smoothing (Ghosh & Narayanan 2011) have been applied to EMA data (see Section 4.3). Neural networks also constitute a natural approach to modeling the complex relationship between acoustic signal and articulatory configuration (Rahim et al. 1991). For instance, Richmond (2006) used a mixture density network for speech inversion—essentially a mixture model whose parameters are predicted from the acoustic input through a multilayer perceptron (see section 13.2 of Murphy 2022). More recent approaches use deep learning methods to estimate EMA trajectories, X-ray microbeam trajectories, or even magnetic resonance imaging (MRI) images directly (Seneviratne et al. 2019, Siriwardena et al. 2022).

4.5. Sound Change

Interesting statistical methods arise in historical phonetics, the study of how pronunciations change over time (Beddor 2023), and comparative phonetics, the study of how the sounds of languages are related. Sound changes are often rooted in the exaggeration of inherent phonetic biases

in articulation and perception. While they often arise due to such factors internal to a language, sound changes can also be caused by social influences, contact between languages, or long-term separation of related languages, sometimes as a result of major political changes (e.g., conquests) or geographical factors (e.g., migration). Sound changes can have different effects on the phonological system of a language, such as creating, merging (see Section 4.1.3), or rearranging phonemic contrasts[4] that govern which words have distinct pronunciations. Sound changes are usually regular, meaning that they affect all words containing the relevant sounds or sound patterns, but they can be irregular or sporadic, affecting only some words or forms (e.g., very high-frequency words or phrases, or words borrowed from other languages). An example of borrowing is the sound /ʒ/, which was adopted into English from French via words such as "pleasure" and "rouge" during a period of intense contact between the languages. An example of regular sound change is the pronunciation of /t/ (here, as is customary, we give phonemic transcripts in slashes) in American English after stressed vowels (e.g., "b**u**tter," "**a**tom") as a "tap" closer to [d], where most British English dialects retain [t].

Phonetic evidence has been central to the study of sound change since the nineteenth century. Work in sociolinguistics (the subfield of linguistics focusing on language in society; e.g., Labov 2001, Eckert 2012) has focused on how sound changes spread through a community and how they become integrated into a language's phonological system, emphasizing the interplay of social factors and speech perception/production. Work in phonetics (e.g., Ohala 1993, Garrett & Johnson 2013, Beddor 2023) has focused on the interaction of speech perception and production to explain why some sound changes occur much more frequently than others. An open question in sound change is the actuation problem: explaining why sound change occurs at some times but not others (Baker et al. 2011, Yu 2023), which has been the subject of computational modeling work (Niyogi 2006, Kirby & Sonderegger 2013).

Traditionally, sound change has been studied using symbolic (i.e., alphabetic) representation of the normative way these sounds were pronounced in lists of cognate words (e.g., Jäger 2019). A notable development has been the interpretation of sound changes in terms of vocal tract gestures (Browman & Goldstein 1991, Carignan et al. 2021), using the models discussed in Section 4.3. However, such approaches have not yet been applied systematically to larger speech corpora.

More recently, there has been growing interest in modeling sound change using speech recordings in historical phonetics, thus incorporating the variability of pronunciations within a language or dialect. This idea was first suggested by the Functional Phylogenies Group (2012), which considered speech corpora from modern languages as data observed at the leaves of an evolutionary tree. This approach postulates that sound change can be described by a continuous process in some sound representation space (such as spectrograms or MFCCs), with the relationships between languages represented by an evolutionary tree and with nodes representing clear separation of the direction of changes between two subbranches of the tree. In this way, tree structure learned from textual analysis or historical evidence can be incorporated into models with the hope of inferring past sound change, for example, through Gaussian process regression (Rasmussen & Williams 2005), where the kernel function is based on the tree structure (Functional Phylogenies Group

[4]Two sounds are phonemes that contrast in a language if two words differing only in these sounds differ in meaning, and speakers agree that they sound different. For example, we know /i/ and /o/ contrast in modern English because "meet" and "moat" are different words and any English speaker would say they are pronounced differently. Phonemes are language-specific. For example, the distinction between aspirated [t^h], as in careful, standard pronunciations of "water," and the glottal stop [ʔ], as in colloquial, informal pronunciations of "water," does not constitute a phonemic contrast in English, while it does so in Hebrew, because there are word pairs in which those two consonants occur in distinct words, such as [t^hor] "queue" versus [ʔor] "light."

2012). Also, information from modern speech corpora can be used to infer or verify hypothesized language tree structures (e.g., Shiers et al. 2017). One important question for this approach is: What sound object characterizes language pronunciation? Average sound representations (spectrograms or MFCCs, for example), as well as covariability of these objects, have been considered (Pigoli et al. 2018). Coleman et al. (2015) use matrices of LPC parameters of whole words in order to model sound changes (e.g., Latin [tres] → Portuguese [treiʃ]), Pigoli et al. (2018) use FFT spectrograms of whole words as data objects, and Hudson et al. (2024) use matrices of MFCC parameters as data objects representing word pronunciations. However, the evolutionary models considered until now are too simple to be realistic and suffer from the tension between the continuous representation of the sound change and the existence of apparently abrupt changes in the language (Pulleyblank 1978). Saltatory models based on Brownian motion, similar to those used in evolutionary biology, have been suggested (Hudson et al. 2024), but this remains an underexplored topic.

Computational models have been proposed to reproduce sound change, based on dynamical systems (Niyogi 2006, Sonderegger & Niyogi 2010, Kirby & Sonderegger 2013), where the way parents transmit their pronunciation to children is modeled through speaker–listener exemplar-based probabilistic models (Bybee 2001, Todd et al. 2019) and agent-based models[5] (de Boer 2001, Stevens et al. 2019, Gubian et al. 2023). The latter simulate sound change in a population by modeling a population of speakers as different probability distribution–valued stochastic processes (agents) who influence one another through complex interactions.

4.6. Machine Learning and Speech Technology in Phonetics Research

There have been continual and mutual exchanges of knowledge between the fields of phonetics and speech technology. Two representative speech technology tasks are ASR and text-to-speech synthesis (TTS). The goal of ASR is to transcribe spoken words as faithfully as possible. The goal of TTS is to produce natural-sounding speech from textual input.

Until the early 2010s, speech technology was based largely on supervised learning. The problem of ASR was formalized as the maximization of P(word sequence|speech signal), which by Bayes' theorem can be decomposed as the maximization of the product of P(speech signal|word sequence) (i.e., an acoustic model) and P(word sequence) (i.e., a language model) (Rabiner & Juang 1993). The acoustic model was usually implemented as an HMM over sequences of phonetic units, for instance, each HMM state mapping to a portion of a vowel or consonant, combined with a phonetic dictionary bridging between the (preprocessed) acoustic input and the sequence of written words. A similar, often richer, architecture was the base for TTS, where the role of HMMs was to generate sequences of speech parameters (fundamental frequency, vowel formants, etc.) based on the input text, which would then be turned into an acoustic signal by a vocoder (Zen et al. 2009, Erro et al. 2014). All these systems required accurately labeled speech corpora as training material. ASR required orthographically transcribed speech corpora associated with a phonetic dictionary as a minimum, while better training materials consist of manually generated or manually checked phonetic transcriptions (e.g., the TIMIT corpus, first released in 1988 and still in use; Garofolo et al. 1993).

[5]Agent-based models are computational models used to simulate and analyze how linguistic changes in pronunciation (sound changes) occur and spread within a population. In these models, individuals (agents) interact according to specified rules, and these interactions can lead to the adoption or evolution of new speech patterns. Agent-based models allow researchers to observe how certain factors, like social network structures or frequency of interaction, can influence the process of sound change over time. This approach provides a dynamic way to study language evolution, capturing complexities that other models might miss.

Supplemental Material >

An important by-product of phonetic HMM-based ASR is the possibility of estimating the location of phonetic boundaries in the input speech signal, using the fact that the maximization of P(speech signal|word sequence) is carried out by applying the Viterbi (1967) algorithm to the concatenation of the hidden states corresponding to the hypothesized phonetic sequence. The transitions between the hidden states obtained from the Viterbi algorithm are used to estimate phonetic boundaries. A special case known as forced alignment consists in retrieving the timings of these boundaries for a known phonetic sequence, such as the timing of the transitions between [k] and [æ] and between [æ] and [t] in the known sequence [kæt] ("cat"). In this case, the Viterbi algorithm is run only once on the known hidden state sequence, rather than several times on many candidate sequences. Although of limited commercial interest, forced alignment represents a major contribution from speech technology to phonetic sciences (Yuan & Liberman 2008, Reddy & Stanford 2015, Kisler et al. 2017, McAuliffe et al. 2017). In particular, it facilitated the growth of corpus phonetics (Harrington 2010, Liberman 2019), marking the transition from traditional phonetics, where controlled experiments are performed to collect limited amounts of speech data, to the analysis of observational data from speech corpora hundreds or thousands of hours in size (see the **Supplemental Appendix**).

The past decade or so has been marked by the advent of deep neural networks (DNNs), which have revolutionized speech technology (Hinton et al. 2012). Two major innovations characterize the state of the art of ASR, TTS and related tasks. One is the end-to-end approach, where system components designed on the basis of domain knowledge, like the phonetically informed acoustic models in ASRs described above, are replaced by a general-purpose computational architecture (a DNN) trained using input–output associations, for example, from sounds to words in the case of ASR. This has decreased the need for phonetically annotated corpora and phonetic dictionaries in the design of speech technology systems, particularly for high-resource languages (e.g., English, Mandarin), though significant challenges remain in ASR for low-resource languages, which are of significant interest for phonetics. The second paradigm shift is the introduction of self-supervised representation learning (SSL) (Mohamed et al. 2022), where a DNN learns to predict its input or to reconstruct its randomly masked input. In the case of ASR, SSL is employed to learn context-rich numerical representations from untranscribed speech datasets of unprecedented size (decades). The pretrained DNN is then fine-tuned, supervised by an orthographically transcribed (small) speech corpus.

5. CONCLUSIONS AND OPEN PROBLEMS IN PHONETICS RESEARCH

Phonetics abounds with data, be they acoustic speech recordings (Sections 3–4.1), neuroimaging data (Section 4.2), or articulatory data (Section 4.3). Landmark shifts in the statistical modeling of such data include the introduction of random effects, the modeling of curve data, and machine learning methods such as forced alignment. Some open statistical modeling questions in phonetics are the following.

Many studies are based on the fundamental frequency f_0 and the formant frequencies. The accuracy of those measurements cannot be taken for granted, so they are manually checked in many studies. However, the perspective taken is mostly a signal-processing one, and a more statistical approach is lacking: What underlying parameter is being estimated when computing f_0 and formant frequencies? This presents a significant challenge, because unlike the relatively straightforward problem of estimating (say) the expectation of a random variable X from an independent sample $x_1, \ldots, x_n$, formulating a probabilistic model for a speech signal with specific values of f_0 and formant frequencies as its parameters is far more complex.

A limitation of the GAMM approach to phonetics data is that variable selection may be challenging. In phonetics, the need for better guidance on these procedures has been expressed (e.g.,

Sóskuthy 2021). One possible solution is the use of penalization (shrinkage) methods, such as the lasso method (Hastie et al. 2009), which have proved effective both for generalized additive models (Marra & Wood 2011, Bai et al. 2022) and for the selection and shrinkage of fixed (Schelldorfer et al. 2014) and random (Pan & Huang 2014) effects in generalized linear mixed-effects models. To the best of our knowledge, the application of these methods to GAMMs has been limited to the case where only the fixed effects are modeled by smooth functions (Lai et al. 2012), though extending them to include smooth functions as random effects would be a useful development. Note, however, that although most of the important acoustic variables of interest change smoothly, not all do: Going beyond smooth functions would also be of interest.

A big current challenge is to scale up speech models to take into account the growing amount of data now easily accessible online, such as audio recordings and audio-in-video of spontaneous and continuous speech from many diverse speakers. Indeed, while forced alignment makes it possible to locate and extract hundreds of thousands of formant trajectories from vowel tokens in a large corpus, modeling such a quantity of multidimensional trajectories with GAMMs or other FDA techniques remains prohibitively expensive.

Extending current models to account for higher-level phonetic and linguistic aspects of speech—such as grammar, semantics, tone of voice (e.g., sarcasm, irony), discourse, and conversational interaction—is of interest. This would shift the current focus on word pronunciations to how features of sound are used to encode higher-level linguistic and communicative structure. Other challenges lie in modeling data that are quite easy to obtain in fairly large quantities—such as ultrasound tongue imaging—but difficult to process because doing so involves time series of very noisy images.

Modeling speech in the brain using technology that translates neural activity (measured, for instance, via electrode arrays) into speech is also a very active research area (e.g., Anumanchipalli et al. 2019).

The recent developments in SSL for speech data have caught the attention of researchers interested in discovering whether the representations learned from the acoustic signal (without supervision) bear any relation to concepts from phonetics, phonology, and so forth. Are phonetic and speaker information distinct in the SSL representation space? Do explicit representations of phones emerge? Is there any resemblance to the cognitive representations formed in the human brain? These hard questions come with a number of statistical challenges, as the analysis often starts with millions of large-dimensional vectors (Beguš 2021, Pasad et al. 2021, Mohamed et al. 2022, tom Dieck et al. 2022).

DISCLOSURE STATEMENT

The authors are unaware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Stefano Coretta, Almond Stöcker, Olivier Renaud, Dominique-Laurent Couturier, Josiane Riverin-Coutlée and Meghan Clayards for helpful discussion and suggestions. We also thank the Editor and reviewer for comments that improved the quality of the article. J.C.'s contribution to this article was supported by a Leverhulme Trust Major Research Fellowship.

LITERATURE CITED

Anumanchipalli GK, Chartier J, Chang EF. 2019. Speech synthesis from neural decoding of spoken sentences. *Nature* 568(7753):493–98

Aston JAD, Chiou JM, Evans JP. 2010. Linguistic pitch analysis using functional principal component mixed effect models. *J. R. Stat. Soc. C* 59(2):297–317

Atal B, Chang JJ, Mathews MV, Tukey JW. 1978. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *J. Acoust. Soc. Am.* 63(5):1535–53

Atal B, Schroeder MR. 1978. Predictive coding of speech signals and subjective error criteria. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 573–76. Piscataway, NJ: IEEE

Baayen RH. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge, UK: Cambridge Univ. Press

Baayen RH, Davidson D, Bates D. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59(4):390–412

Bai R, Moran GE, Antonelli JL, Chen Y, Boland MR. 2022. Spike-and-slab group lassos for grouped regression and sparse generalized additive models. *J. Am. Stat. Assoc.* 117(537):184–97

Baker A, Archangeli D, Mielke J. 2011. Variability in American English s-retraction suggests a solution to the actuation problem. *Lang. Var. Change* 23(3):347–74

Barr DJ. 2008. Analyzing 'visual world' eyetracking data using multilevel logistic regression. *J. Mem. Lang.* 59(4):457–74

Barreda S. 2023. phonTools: functions for phonetics in R. *R Package*, version 0.2-2.2. **https://cran.r-project.org/web/packages/phonTools/**

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67(1):1–48

Beautemps D, Badin P, Laboissière R. 1995. Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: a new model for vowels and fricative consonants based on experimental data. *Speech Commun.* 16:27–47

Beddor PS. 2023. Advancements of phonetics in the 21st century: theoretical and empirical issues in the phonetics of sound change. *J. Phon.* 98:101228

Beguš G. 2021. ciwGAN and fiwGAN: encoding information in acoustic data to model lexical learning with generative adversarial networks. *Neural Netw.* 139:305–25

Boersma P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proc. Inst. Phon. Sci.* 17:97–110

Boersma P, Weenink D. 2023. Praat: doing phonetics by computer. *Phonetics Software*, version 6.3.17. **http://www.praat.org**

Bogert BP, Healy MJR, Tukey JW. 1963. The quefrency alanysis of time series for echoes: cepstrum, pseudoautocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the Symposium in Time Series Analysis, 1962*, ed. M Rosenblatt, pp. 209–43. New York: Wiley

Brand J, Hay J, Clark L, Watson K, Sóskuthy M. 2021. Systematic co-variation of monophthongs across speakers of New Zealand English. *J. Phon.* 88:101096

Browman CP, Goldstein L. 1991. Gestural structures: distinctiveness, phonological processes, and historical change. In *Modularity and the Motor Theory of Speech Perception*, ed. IG Mattingly, M Studdert-Kennedy, pp. 313–38. New York: Psychology

Bürkner PC. 2017. brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80(1):1–28

Bybee J. 2001. *Phonology and Language Use*. Cambridge, UK: Cambridge Univ. Press

Carignan C, Coretta S, Frahm J, Harrington J, Hoole P, et al. 2021. Planting the seed for sound change: evidence from real-time MRI of velum kinematics in German. *Language* 97(2):333–64

Carignan C, Hoole P, Kunay E, Pouplier M, Joseph A, et al. 2020. Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI. *Lab. Phonol.* 11(1):2

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, et al. 2016. Stan: a probabilistic programming language. *J. Stat. Softw.* 76:1–32

Cederbaum J, Pouplier M, Hoole P, Greven S. 2016. Functional linear mixed models for irregularly or sparsely sampled data. *Stat. Model.* 16(1):67–88

Cedergren HJ, Sankoff D. 1974. Variable rules: performance as a statistical reflection of competence. *Language* 50(2):333–55

Cho T, Whalen D, Docherty G. 2019. Voice onset time and beyond: exploring laryngeal contrast in 19 languages. *J. Phon.* 72:52–65

Chodroff E, Wilson C. 2022. Uniformity in phonetic realization: evidence from sibilant place of articulation in American English. *Language* 98(2):250–89

Chuang YY, Fon J, Papakyritsis I, Baayen H. 2021. Analyzing phonetic data with generalized additive mixed models. In *Manual of Clinical Phonetics*, ed. MJ Ball, pp. 108–38. Abingdon-on-Thames, UK: Routledge

Coleman J. 2005. *Introducing Speech and Language Processing*. Cambridge, UK: Cambridge Univ. Press

Coleman J, Aston JAD, Pigoli D. 2015. *Reconstructing the sounds of words from the past*. Paper presented at the 18th International Congress of Phonetic Sciences, Glasgow, UK, Aug. 10–14

Cooley JW, Tukey JW. 1965. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* 19(90):297–301

Davidson L. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *J. Acoust. Soc. Am.* 120(1):407–15

de Boer B. 2001. *The Origins of Vowel Systems*. Oxford, UK: Oxford Univ. Press

Delattre PC, Liberman AM, Cooper FS. 1955. Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.* 27(4):769–73

Eager C, Roy J. 2017. Mixed effects models are sometimes terrible. arXiv:1701.04858 [stat.AP]

Eckert P. 2012. Three waves of variation study: the emergence of meaning in the study of sociolinguistic variation. *Annu. Rev. Anthropol.* 41:87–100

Erro D, Sainz I, Navas E, Hernaez I. 2014. Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE J. Sel. Top. Signal Process.* 8(2):184–94

Evans J, Chu M, Aston JA, Su C. 2010. Linguistic and human effects on F_0 in a tonal dialect of Qiang. *Phonetica* 67(1/2):82–99

Fant CGM. 1960. *Acoustic Theory of Speech Production*. The Hague, Neth.: Mouton

Fant CGM. 1972. Vocal tract wall effects, losses, and resonance bandwidths. *Speech Transm. Lab. Q. Prog. Status Rep.* 13(2/3):28–52

Fant CGM. 1980. The relations between area functions and the acoustic signal. *Phonetica* 37:55–86

Frossard J, Renaud O. 2022. The cluster depth tests: toward point-wise strong control of the family-wise error rate in massively univariate tests with application to M/EEG. *NeuroImage* 247:118824

Fruehwald J. 2017. Generations, lifespans, and the zeitgeist. *Lang. Var. Change* 29(1):1–27

Functional Phylogenies Group. 2012. Phylogenetic inference for function-valued traits: speech sound evolution. *Trends Ecol. Evol.* 27(3):160–66

Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, et al. 1993. *TIMIT acoustic-phonetic continuous speech corpus LDC93S1*. Speech Corpus Recording Database, Linguist. Data Consort., Philadelphia. **https://doi.org/10.35111/17gk-bn40**

Garrett A, Johnson K. 2013. Phonetic bias in sound change. In *Origins of Sound Change: Approaches to Phonologization*, ed. ACL Yu, pp. 51–97. Oxford, UK: Oxford Univ. Press

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013. *Bayesian Data Analysis*. Boca Raton, FL: CRC. 3rd ed.

Ghosh PK, Narayanan SS. 2011. A subject-independent acoustic-to-articulatory inversion. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 4624–27. Piscataway, NJ: IEEE

Gold B, Morgan N, Ellis D. 2011. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. New York: Wiley. 2nd ed.

Grabe E, Kochanski G, Coleman J. 2007. Connecting intonation labels to mathematical descriptions of fundamental frequency. *Lang. Speech* 50(3):281–310

Groppe DM, Urbach TP, Kutas M. 2011. Mass univariate analysis of event-related brain potentials/fields. I: A critical tutorial review. *Psychophysiology* 48(12):1711–25

Gubian M, Cronenberg J, Harrington J. 2023. Phonetic and phonological sound changes in an agent-based model. *Speech Commun.* 147:93–115

Gubian M, Torreira F, Boves L. 2015. Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. *J. Phon.* 49:16–40

Guo W. 2002. Functional mixed effects models. *Biometrics* 58(1):121–28
Hadjipantelis PZ, Aston JA, Müller HG, Evans JP. 2015. Unifying amplitude and phase analysis: a compositional data approach to functional multivariate mixed-effects modeling of Mandarin Chinese. *J. Am. Stat. Assoc.* 110(510):545–59
Hagiwara R. 1997. Dialect variation and formant frequency: the American English vowels revisited. *J. Acoust. Soc. Am.* 102(1):655–58
Harrington J. 2010. *Phonetic Analysis of Speech Corpora*. Hoboken, NJ: Wiley-Blackwell
Harshman R, Ladefoged P, Goldstein L. 1977. Factor analysis of tongue shapes. *J. Acoust. Soc. Am.* 62(3):693–707
Hastie T, Tibshirani R, Friedman JH. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. 2nd ed.
Hillenbrand J, Getty LA, Clark MJ, Wheeler K. 1995. Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97(5):3099–111
Hinton G, Deng L, Yu D, Dahl GE, Mohamed A, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* 29(6):82–97
Hiroya S, Honda M. 2004. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Trans. Speech Audio Process.* 12(2):175–85
Hudson T, Wei J, Coleman J. 2024. Using acoustic-phonetic simulations to model historical sound change. *Diachronica* 41(3):355–78
Jäger G. 2019. Computational historical linguistics. *Theor. Linguist.* 45(3–4):151–82
Johnson K. 2012. *Acoustic and Auditory Phonetics*. Hoboken, NJ: Wiley-Blackwell. 3rd ed.
Johnson K, Sjerps MJ. 2021. Speaker normalization in speech perception. In *The Handbook of Speech Perception*, ed. JS Pardo, LC Nygaard, RE Remez, DB Pisoni, pp. 145–76. New York: Wiley. 2nd ed.
Kaan E. 2007. Event-related potentials and language processing: a brief overview. *Lang. Linguist. Compass* 1(6):571–91
Kaipio J, Somersalo E. 2005. *Statistical and Computational Inverse Problems*. New York: Springer
Kawahara H, Morise M, Takahashi T, Nisimura R, Irino T, Banno H. 2008. Tandem-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3933–36. Piscataway, NJ: IEEE
Kendall T, Pharao N, Stuart-Smith J, Vaughn C. 2023. Advancements of phonetics in the 21st century: theoretical issues in sociophonetics. *J. Phon.* 98:101226
Kirby J, Sonderegger M. 2013. A model of population dynamics applied to phonetic change. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pp. 776–81. Seattle, WA: Cogn. Sci. Soc.
Kiritani S, Itoh K, Fujimura O. 1975. Tongue-pellet tracking by a computer-controlled X-ray microbeam system. *J. Acoust. Soc. Am.* 57(6):1516–20
Kisler T, Reichel U, Schiel F. 2017. Multilingual processing of speech via web services. *Comput. Speech Lang.* 45:326–47
Klatt DH, Klatt LC. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.* 87(2):820–57
Koenig LL, Lucero JC, Perlman E. 2008. Speech production variability in fricatives of children and adults: results of functional data analysis. *J. Acoust. Soc. Am.* 124(5):3158–70
Kohler KJ. 1987. Categorical pitch perception. In *Proceedings of the XIth International Congress of Phonetic Sciences*, Vol. 5, pp. 331–33. London: Int. Phon. Assoc.
Koshy A, Tavakoli S. 2022. Exploring British accents: modelling the trap–bath split with functional data analysis. *J. R. Stat. Soc. C* 71(4):773–805
Labov W. 2001. *Principles of Linguistic Change*, Vol. 2: *Social Factors*. Malden, MA: Blackwell
Lai RCS, Huang HC, Lee TCM. 2012. Fixed and random effects selection in nonparametric additive mixed models. *Electron. J. Stat.* 6:810–42
Larar JN, Schroeter J, Sondhi MM. 1988. Vector quantization of the articulatory space. *IEEE Trans. Acoust. Speech Signal Process.* 36(12):1812–18
Liberman MY. 2019. Corpus phonetics. *Annu. Rev. Linguist.* 5:91–107

Lim Y, Toutios A, Bliesener Y, Tian Y, Lingala SG, et al. 2021. A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images. *Sci. Data* 8:187

Lisker L, Abramson AS. 1964. A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20(3):384–422

Lisker L, Abramson AS. 1970. The voicing dimension: some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences*, ed. B Hála, M Romportl, P Janota, pp. 563–65. London: Int. Phon. Assoc.

Lobanov BM. 1971. Classification of Russian vowels spoken by different speakers. *J. Acoust. Soc. Am.* 49(2B):606–8

Lucero JC, Koenig LL. 2000. Time normalization of voice signals using functional data analysis. *J. Acoust. Soc. Am.* 108(4):1408–20

Malinen J, Palo P. 2009. Recording speech during MRI: part II. In *Models and Analysis of Vocal Emissions for Biomedical Applications: 6th International Workshop*, ed. C Manfredi, pp. 211–14. Firenze, Italy: Firenze Univ. Press

Marra G, Wood SN. 2011. Practical variable selection for generalized additive models. *Comput. Stat. Data Anal.* 55(7):2372–87

Marron JS, Ramsay JO, Sangalli LM, Srivastava A. 2015. Functional data analysis of amplitude and phase variation. *Stat. Sci.* 30(4):468–84

McAuliffe M, Socolof M, Mihuc S, Wagner M, Sonderegger M. 2017. Montreal forced aligner: trainable text-speech alignment using Kaldi. In *Proceedings of INTERSPEECH 2017*, pp. 498–502. N.p.: Int. Speech Commun. Assoc.

Mermelstein P. 1967. Determination of the vocal-tract shape from measured formant frequencies. *J. Acoust. Soc. Am.* 41(5):1283–94

Mielke J, Carignan C, Thomas E. 2017. The articulatory dynamics of pre-velar and pre-nasal /æ/-raising in English: an ultrasound study. *J. Acoust. Soc. Am.* 142(1):332–49

Mirman D, Dixon JA, Magnuson JS. 2008. Statistical and computational models of the visual world paradigm: growth curves and individual differences. *J. Mem. Lang.* 59(4):475–94

Mohamed A, Lee H, Borgholt L, Havtorn JD, Edin J, et al. 2022. Self-supervised speech representation learning: a review. *IEEE J. Sel. Top. Signal Process.* 16(6):1179–210

Moisik S, Lin H, Esling J. 2014. A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). *J. Int. Phon. Assoc.* 44(1):21–58

Mrayati M, Carré R, Guérin B. 1988. Distinctive regions and modes: a new theory of speech production. *Speech Commun.* 7(3):257–86

Murphy KP. 2022. *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: MIT Press

Murphy KP. 2023. *Probabilistic Machine Learning: Advanced Topics*. Cambridge, MA: MIT Press

Näätänen R, Gaillard AW, Mäntysalo S. 1978. Early selective-attention effect on evoked potential reinterpreted. *Acta Psychol.* 42(4):313–29

Nixon JS, Van Rij J, Mok P, Baayen RH, Chen Y. 2016. The temporal dynamics of perceptual uncertainty: eye movement evidence from Cantonese segment and tone perception. *J. Mem. Lang.* 90:103–25

Niyogi P. 2006. *The Computational Nature of Language Learning and Evolution*. Cambridge, MA: MIT Press

Ohala JJ. 1993. Sound change as nature's speech perception experiment. *Speech Commun.* 13(1/2):155–61

Ohala JJ. 2006. Phonetics: overview. In *Encyclopedia of Language and Linguistics*, ed. K Brown, pp. 468–70. Oxford, UK: Elsevier. 2nd ed.

Ombao H, Lindquist M, Thompson W, Aston J, eds. 2016. *Handbook of Neuroimaging Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC

Pan J, Huang C. 2014. Random effects selection in generalized linear mixed models via shrinkage penalty function. *Stat. Comput.* 24:725–38

Pasad A, Chou JC, Livescu K. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–21. Piscataway, NJ: IEEE

Pigoli D, Hadjipantelis PZ, Coleman JS, Aston JAD. 2018. The statistical analysis of acoustic phonetic data: exploring differences between spoken Romance languages. *J. R. Stat. Soc. C* 67(5):1103–45

Pini A, Spreafico L, Vantini S, Vietti A. 2019. Multi-aspect local inference for functional data: analysis of ultrasound tongue profiles. *J. Multivar. Anal.* 170:162–85
Pisoni DB, Tash J. 1974. Reaction times to comparisons within and across phonetic categories. *Percept. Psychophys.* 15(2):285–90
Pouplier M, Cederbaum J, Hoole P, Marin S, Greven S. 2017. Mixed modeling for irregularly sampled and correlated functional data: speech science applications. *J. Acoust. Soc. Am.* 142(2):935–46
Proudfoot M, Woolrich MW, Nobre AC, Turner MR. 2014. Magnetoencephalography. *Pract. Neurol.* 14(5):336–43
Puggaard-Rode R. 2022. Analyzing time-varying spectral characteristics of speech with function-on-scalar regression. *J. Phon.* 95:101191
Pulleyblank EG. 1978. Abruptness and gradualness in phonological change. In *Linguistic and Literary Studies*, Vol. 3: *Historical and Comparative Linguistics*, ed. MA Jazayery, EC Polomé, W Winter, pp. 181–92. Berlin: De Gruyter
Rabiner L, Juang BH. 1993. *Fundamentals of Speech Recognition*. Hoboken, NJ: Prentice-Hall
Rahim M, Keijn W, Schroeter J, Goodyear C. 1991. Acoustic to articulatory parameter mapping using an assembly of neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 485–88. Piscataway, NJ: IEEE
Ramsay JO, Munhall K, Gracco V, Ostry D. 1996. Functional data analyses of lip motion. *J. Acoust. Soc. Am.* 99:3718–27
Ramsay JO, Silverman BW. 2005. *Functional Data Analysis*. New York: Springer. 2nd ed.
Rasmussen CE, Williams CKI. 2005. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press
Rebernik T, Jacobi J, Jonkers R, Noiray A, Wieling M. 2021. A review of data collection practices using electromagnetic articulography. *Lab. Phonol.* 12(1):6
Reddy S, Stanford JN. 2015. Toward completely automated vowel extraction: introducing DARLA. *Linguist. Vanguard* 1(1):15–28
Renwick MEL, Stanley JA. 2020. Modeling dynamic trajectories of front vowels in the American South. *J. Acoust. Soc. Am.* 147(1):579–95
Renwick MEL, Stanley JA, Forrest J, Glass L. 2023. Boomer peak or Gen X cliff? From SVS to LBMS in Georgia English. *Lang. Var. Change* 35(2):175–97
Richmond K. 2006. A trajectory mixture density network for the acoustic-articulatory inversion mapping. In *Proceedings of INTERSPEECH 2006*, pp. 577–80. N.p.: Int. Speech Commun. Assoc.
Rosner BS, Pickering JB. 1994. *Vowel Perception and Production*. Oxford, UK: Oxford Univ. Press
Schafer RW, Rabiner LR. 1970. System for automatic formant analysis of voiced speech. *J. Acoust. Soc. Am.* 47(2B):634–48
Schelldorfer J, Meier L, Bühlmann P. 2014. GLMMLasso: an algorithm for high-dimensional generalized linear mixed models using ℓ_1-penalization. *J. Comput. Graph. Stat.* 23(2):460–77
Schroeder MR. 1967. Determination of the geometry of the human vocal tract by acoustic measurements. *J. Acoust. Soc. Am.* 41(4B):1002–10
Schroeder MR. 1985. Linear predictive coding of speech: review and current directions. *IEEE Commun. Mag.* 23(8):54–61
Schroeter J, Meyer P, Parthasarathy S. 1990. Evaluation of improved articulatory codebooks and codebook access distance measures. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 393–96. Piscataway, NJ: IEEE
Schroeter J, Sondhi M. 1989. Dynamic programming search of articulatory codebooks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 588–91. Piscataway, NJ: IEEE
Seneviratne N, Sivaraman G, Espy-Wilson C. 2019. Multi-corpus acoustic-to-articulatory speech inversion. In *Proceedings of INTERSPEECH 2019*, pp. 859–63. N.p.: Int. Speech Commun. Assoc.
Shiers N, Aston JA, Smith JQ, Coleman JS. 2017. Gaussian tree constraints applied to acoustic linguistic functional data. *J. Multivar. Anal.* 154:199–215
Shumway RH, Stoffer DS. 2017. *Time Series Analysis and Its Applications: With R Examples*. Cham, Switz.: Springer. 4th ed.
Siriwardena YM, Sivaraman G, Espy-Wilson C. 2022. Acoustic-to-articulatory speech inversion with multi-task learning. In *Proceedings of INTERSPEECH 2022*, pp. 5020–24. N.p.: Int. Speech Commun. Assoc.

Sonderegger M. 2023. *Regression Modeling for Linguistic Data*. Cambridge, MA: MIT Press
Sonderegger M, Niyogi P. 2010. Combining data and mathematical models of language change. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1019–29. Stroudsburg, PA: Assoc. Comput. Linguist.
Sonderegger M, Sóskuthy M. 2024. Advancements of phonetics in the 21st century: quantitative data analysis. PsyArXiv mc6a9. **https://osf.io/preprints/psyarxiv/mc6a9**
Sóskuthy M. 2021. Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *J. Phon.* 84:101017
Stevens KN. 1998. *Acoustic Phonetics*. Cambridge, MA: MIT Press
Stevens M, Harrington J, Schiel F. 2019. Associating the origin and spread of sound change using agent-based modelling applied to /s/-retraction in English. *Glossa J. Gen. Linguist.* 4(1):8
Stone M. 1990. A three-dimensional model of tongue movement based on ultrasound and X-ray microbeam data. *J. Acoust. Soc. Am.* 87(5):2207–17
Tagliamonte SA. 2002. *Analyzing Sociolinguistic Variation*. Cambridge, UK: Cambridge Univ. Press
Talkin D. 1995. A robust algorithm for pitch tracking (RAPT). In *Speech Coding and Synthesis*, ed. WB Kleijn, KK Paliwal, pp. 497–518. Amsterdam: Elsevier
Tanner J, Sonderegger M, Stuart-Smith J, Fruehwald J. 2020. Toward "English" phonetics: variability in the pre-consonantal voicing effect across English dialects and speakers. *Front. Artif. Intel.* 3:38
Tavakoli S, Pigoli D, Aston JAD, Coleman JS. 2019. A spatial modeling approach for linguistic object data: analyzing dialect sound variations across Great Britain. *J. Am. Stat. Assoc.* 114(527):1081–96
Thul R, Conklin K, Barr DJ. 2021. Using GAMMs to model trial-by-trial fluctuations in experimental data: more risks but hardly any benefit. *J. Mem. Lang.* 120:104247
Toda T, Black A, Tokuda K. 2004. Acoustic-to-articulatory inversion mapping with Gaussian mixture model. In *Proceedings of INTERSPEECH 2004*, pp. 1129–32. N.p.: Int. Speech Commun. Assoc.
Todd S, Pierrehumbert JB, Hay J. 2019. Word frequency effects in sound change as a consequence of perceptual asymmetries: an exemplar-based model. *Cognition* 185:1–20
tom Dieck T, Pérez-Toro PA, Arias T, Nöth E, Klumpp P. 2022. Wav2vec behind the scenes: how end2end models learn phonetics. In *Proceedings of INTERSPEECH 2022*, pp. 5130–34. N.p.: Int. Speech Commun. Assoc.
Tomaschek F, Wieling M, Arnold D, Baayen RH. 2013. Word frequency, vowel length and vowel quality in speech production: an EMA study of the importance of experience. In *Proceedings of INTERSPEECH 2013*, pp. 1302–6. N.p.: Int. Speech Commun. Assoc.
Toutios A, Narayanan S. 2016. Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research. *APSIPA Trans. Signal Inform. Proc.* 5:e6
Vasishth S, Nicenboim B, Beckman ME, Li F, Kong EJ. 2018. Bayesian data analysis in the phonetic sciences: a tutorial introduction. *J. Phon.* 71:147–61
Viterbi A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* 13(2):260–69
Voeten CC, Heeringa W, Van De Velde H. 2022. Normalization of nonlinearly time-dynamic vowels. *J. Acoust. Soc. Am.* 152(5):2692–710
Volkmann A, Stöcker A, Scheipl F, Greven S. 2023. Multivariate functional additive mixed models. *Stat. Model.* 23(4):303–26
Westbury J, Turner G, Denbowski J. 1994. *X-ray microbeam speech production database user's handbook*, v. 1.0. Tech. Rep., Waisman Cent. Ment. Retard. Hum. Dev., Univ. Wis., Madison, WI. **https://ubeam.engr.wisc.edu/pdf/ubdbman.pdf**
Wieling M. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *J. Phon.* 70:86–116
Wieling M, Tomaschek F, Arnold D, Tiede M, Bröker F, et al. 2016. Investigating dialectal differences using articulography. *J. Phon.* 59:122–43
Winter B. 2019. *Statistics for Linguists: An Introduction Using R*. London: Routledge
Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. B* 73(1):3–36

Wood SN. 2017. *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall/CRC. 2nd ed.

Yu AC. 2023. The actuation problem. *Annu. Rev. Linguist.* 9:215–31

Yuan J, Liberman M. 2008. Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* 123(Suppl. 5):3878 (Abstr.)

Zen H, Tokuda K, Black AW. 2009. Statistical parametric speech synthesis. *Speech Commun*. 51(11):1039–64