ANNUAL REVIEWS

Annual Review of Linguistics

Corpus Phonetics

Mark Y. Liberman

Department of Linguistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; email: myl@cis.upenn.edu

Annu. Rev. Linguist. 2019. 5:91–107

First published as a Review in Advance on August 22, 2018

The *Annual Review of Linguistics* is online at linguist.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-011516-033830

Keywords

linguistics, phonetics, speech analysis, reproducibility

Abstract

Semiautomatic analysis of digital speech collections is transforming the science of phonetics. Convenient search and analysis of large published bodies of recordings, transcripts, metadata, and annotations—up to three or four orders of magnitude larger than a few decades ago—have created a trend towards "corpus phonetics," whose benefits include greatly increased researcher productivity, better coverage of variation in speech patterns, and crucial support for reproducibility. The results of this work include insights into theoretical questions at all levels of linguistic analysis, along with applications in fields as diverse as psychology, medicine, and poetics, as well as within phonetics itself. Remaining challenges include still-limited access to the necessary skills and a lack of consistent standards. These changes coincide with the broader Open Data movement, but future solutions will also need to include more constrained forms of publication motivated by valid concerns for privacy, confidentiality, and intellectual property.

1. OVERVIEW

Phonetics, the scientific study of speech sounds, has been transformed in recent years due to the possibility of creating very large digital collections of transcribed and aligned audio recordings. Additional factors are shared with other empirical sciences; these include cost–performance improvements in digital hardware and new software for organizing and analyzing large data sets, specifically, research in phonetics from data sets created and published to support research in speech technology, from the use of speech recognition technology to help analyze acoustic data sets, and from the role of digital media in creating large volumes of "found data" in the form of digital audio and video recordings, often with available texts or transcripts. The result is a trend toward corpus phonetics, where research is based on these large published collections of recordings, transcripts, metadata, and annotations.

Data set size is important in phonetic science, because a very large number of varied factors influence speech sounds in systematic ways. These factors include not only language and dialect but also culture and style; not only phonological, lexical, and syntactic context but also discourse and conversational structure; not only the form and content of a linguistic message but also the process of creating and producing it; not only individual characteristics but also individuals' attitudes toward the subject matter and the communicative context; and not only individuals' communicative intent but also their physiological state and their level of vocal effort.

These sources of variation matter even in studies of the basic nature of consonants and vowels in a given language: College students reading citation forms in a sound booth yield very different measurements from people of all sorts conversing, discussing, orating, or even reading out loud in real life. And appropriately varied samples are obviously essential as a basis for phonetic studies of dialect, style, phrasing, conversation, attitude, frequency and recency effects, and so on. As a result, controlled elicitation across a relevant range of factors already requires a large data set; and if we want to base our research on more naturalistic, "ecologically valid" materials, we will need to sample from a much larger collection. Therefore, as it becomes easier to collect and analyze large quantities of speech data, or to access existing multimedia archives, phoneticians have naturally responded by creating and using larger and larger data sets, and by taking increasing advantage of existing material. And the same technological changes make digital archiving and publication of phonetic data sets increasingly easy, shortening the cycle of scientific progress by facilitating replication and extension of results. These developments have elevated corpus phonetics from a marginal position to an increasingly central one.

Phonetics can be conveniently divided into subfields along two dimensions: the types of data studied and the goals of the investigation. Along the first dimension, the major division is between acoustic phonetics, based on the study of speech sounds, and articulatory phonetics, based on the study of the vocal-tract gestures involved in creating those sounds. The second dimension spans the wide range of topics related in some way to human speech. A key goal is to understand the ways in which phonological categories and structures—the representations that define words and phrases—are realized in gestures and sounds. Some researchers are interested in the psychology, neurology, and developmental trajectory of speech production and perception, or in the phonetic projection of interpersonal interaction, rhetorical structure, attitude, and emotion. For other researchers, the focus is on phonetic correlates of clinical categories. An especially important area is the study of phonetic variation in general and its role in sound change, and its relationships to geography and demography.

Along the first dimension, the trend towards corpus phonetics is still largely confined to the acoustic side and to the study of speech production, with articulatory and perceptual research lagging behind. This is due to the nature of the underlying technologies—speech recordings are

simply easier and less expensive to create. In terms of the research topics on the second dimension, corpus-based methods are increasingly used across the board. The trend is especially strong in sociolinguistics (e.g., Kendall 2013) and psycholinguistics (e.g., Meyer et al. 2016), but similar effects can be found as far afield as literary studies (e.g., MacArthur 2016) and political science (e.g., Kirkham & Moore 2016). Remaining challenges include the significant differences between disciplines in access to the necessary skills and tools, a lack of consistent standards, and the fact that the broader Open Data movement (Kitchin 2014) has been more widely accepted in some disciplines than in others. Resistance is based to some extent on cultural differences in proprietary attitudes toward data, but also to some extent on valid concerns for privacy, confidentiality, and intellectual property, where the development of appropriate solutions remains a work in progress.

In the remainder of this review, I discuss the details of the process that has brought the field from the research paradigms of 50 and 100 years ago to its current state. After discussing recent developments and current issues, I close by describing the path that we can expect the field to follow over the next dozen years or so.

2. THE HISTORY

The Oxford English Dictionary (OED) defines *corpus*, in the relevant sense, as "[t]he body of written or spoken material upon which a linguistic analysis is based." Interestingly, the OED's earliest citation for this sense is a quotation from Allen (1956, p. 128) in which *corpus* refers to a collection of tape recordings: "The analysis here presented is based on the speech of a single informant . . . and in particular upon a corpus of material, of which a large proportion was narrative, derived from approximately 100 hours of listening."

Allen's choice of the word *corpus* no doubt took as its model the tradition that began with Livy's phrase *corpus omnis Romani iuris*, "the body of all Roman law," in Book 3.34 of his work *Ab urbe condita*. And although Allen was apparently the first to apply *corpus* to linguistic data, the idea of linguistic investigation based on a predefined body of material is even older, exemplified by Pāṇini's account of Vedic Sanskrit from the sixth century BCE. Over the intervening millennia, successive waves of philologists have focused their analytic attention on text collections of all kinds. Much of this analysis was lexicographical or morphosyntactic, but phonetics was part of the enterprise as well. Nineteenth- and early-twentieth-century philologists used clues such as spelling variation, verse scansion, and the transformation of loan words to speculate about the pronunciation of dead languages. Other researchers organized systematic dialect surveys to try to characterize pronunciation variation in living languages.

2.1. The First Wave: Analog Recording and Analysis

The invention of mechanical recording devices in the late nineteenth century led to several attempts at systematic collection and phonetic analysis of speech data. For example, Jones (1909) based an analysis of "intonation curves" on eight commercially available gramophone records: three English, three French, and two German. Ferdinand Brunot founded the Archives de la Parole at the Sorbonne in 1911 as a repository for his collections of speech recordings, "la première pierre d'un Institut de phonétique que l'université de Paris souhaite mettre en place" (Cordereix 2014, p. 5). And mechanical instruments like the phonautograph (Ellis 1874) allowed some acoustic properties of speech sounds to be measured. However, these efforts were necessarily limited in size and scope. The quantity and quality of the recordings were relatively low, and the phonetic measurement instruments were relatively crude and, in most cases, could not even be used to analyze recorded speech.

Over the next century, a series of technological and social innovations allowed researchers to more closely approach the goals that we can clearly perceive behind research like that of Jones and Brunot. Two of these innovations came in the middle of the twentieth century: the tape recorder (Leslie & Snyder 2010) and the sound spectrograph (Koenig et al. 1946).

The research that Allen described in 1956 (on what he termed the Abaza verbal complex) could not have been carried out in the same way even a few years earlier, because the reel-to-reel tape recorder that he used to create and access his "corpus" was not available until the early 1950s. Audiotape allowed researchers to conveniently record and store significant quantities of speech with relatively high fidelity, and to listen to desired selections (at the cost of considerable time and effort to cue up the recordings, or excise and combine samples by physical cutting and splicing of tape segments). But like other large-scale speech corpus studies during the analog years, Allen's phonetic analysis uses traditionally impressionistic methods. For example, he notes that the "verbal complex" is "defined syntagmatically and phonologically by falling clause-intonation (of which the principal exponent is falling pitch on the last stressed syllable of the clause)" (Allen 1956, p. 131), and presents a dozen schematic pitch contours, created by the same impressionistic process used by Jones (1909).

However, another invention that became available at about the same time as the tape recorder—the sound spectrograph—allowed the exact measurement of phonetic characteristics such as durations, formant frequencies, and pitch contours. And crucially, this instrument could take input from previously recorded audiotapes. Along with advances in signal processing and physical modeling of vocal-tract acoustics, such as formant theory (Chiba & Kajiyama 1941, Fant 1956), the tape recorder and the sound spectrograph together created a first wave of transformation in the field of phonetics, shifting its center of gravity from the humanities to the natural and psychological sciences. Instrumental phonetics gradually superseded impressionistic phonetics, although instrumental measurements still required painstaking human labor—several minutes per measurement or more—and there was still no convenient way to share and reuse collections of speech recordings. As a result, instrumental phonetic analyses were often limited to a handful of illustrative examples, and even the most ambitious projects in this era were based on only a few minutes of speech.

In this first wave of corpus phonetics research, the speech material itself was retained by the researchers, or at best perhaps deposited in some library or archive. In some cases, this was for good practical reasons—Allen's 100 hours of Abaza would have required replication and delivery of a large crate of tapes or audio disks. (As far as I know, this collection was never archived, but was simply lost after Allen's death in 2004.) But in general, the scientific culture of the time did not generally favor publication of data, and so even for the few minutes of speech analyzed in early instrumental studies, only tabulations or statistical summaries were published.

A well-known early example was a study by Peterson & Barney (1952, pp. 176–77), in which "[a] total of 76 speakers, including 33 men, 28 women, and 15 children, each recorded two lists of 10 words, making a total of 1520 recorded words." Durations and formant frequencies were measured from spectrograms, and spliced copies of random samples of these recordings were also perceptually classified by 70 listeners. This research was done at AT&T Bell Laboratories, and aimed at technological as well as scientific goals—until 1980 or so, there was generally close cooperation between phoneticians and speech engineers, with many joint publications and considerable overlap in training and professional development.

Peterson & Barney (1952) presented a summary of perceptual results and a table of average fundamental and formant frequencies for each of 10 vowels in each of three speaker groups (men, women, and children). Nearly 40 years later, Watrous (1991) located printouts of the original data and rescued a full table of all 1,520 vowel (formant and fundamental frequency) measurements

from all 76 speakers; but he found that the audio recordings had been lost. If Peterson and Barney had published a tape or disk containing their 1,520 recorded phrases, we would be happy today to have it; but it would have been of little use to their contemporaries, since there was no way to spare other researchers the complex physical process of finding words, syllables, and segments in the recorded sound stream, or the several hundred hours of labor required to make paper spectrograms, and to measure and record relevant fundamental frequency and formant values.

During this predigital period, much larger bodies of recordings, though without instrumental analysis, were created by various dialect atlas projects. These projects began in earlier decades with lexical checklists and impressionistic pronunciation surveys, but once tape recorders became available around 1950, their standard practice began to include interview recordings. The *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) includes 1,162 interviews collected between 1933 and 1974, with records prior to 1950 "transcribed by the original fieldworker during the interview," while records after 1950 "were transcribed from tape" (Kretzschmar 1993).

Over the years, there have been more than a dozen such projects in the United States alone (see the Linguistic Atlas Projects website at the University of Georgia at **http://us.english.uga.edu/cgi-bin/lapsite.fcgi/** or Kretzschmar 2003 for information about 10 of them), and there have been many other dialect atlas projects around the world. Most of the recordings and other materials from these surveys have been preserved. Some of these materials have recently been digitized, and in some cases the digital version (or a portion of it) has been published. Note that there was relatively little interchange during this period between the dialectologists on one hand and the phoneticians and speech engineers on the other hand. Perhaps for that reason, the phonetic analysis of dialect atlas collections during this period mainly took the form of compilations of impressionistic transcriptions, so the use of this material in instrumental phonetic research remains largely an opportunity for the future.

A related but distinct tradition is the accumulation of sociolinguistic interviews. Labov (1963) famously explored "the social motivation of a sound change" through the analysis of tape-recorded interviews with 69 residents of Martha's Vineyard. His conclusions were based partly on impressionistic analysis of these recordings and partly on 80 formant measurements. Through the next half-century, the growing field of quantitative sociolinguistics generally followed the same pattern, with a collection of taped interviews subject to some combination of impressionistic coding and instrumental measurement of a relatively small number of scattered tokens, typically between a few hundred and a few thousand in any given publication. As with the dialect atlas recordings, most sociolinguists' interview tapes have been preserved, along with associated metadata. And again, many of these archives have recently been digitized, and have been or will be published after appropriate anonymization. Examples include the Sociolinguistics Archive and Analysis Project at North Carolina State University (Kendall 2007) and the thousands of hours of sociolinguistics interviews in the Philadelphia Neighborhood Corpus (Labov et al. 2013).

In the domain of instrumental phonetics and speech technology, most nonsociolinguistic research in the 1960s and 1970s was based on the study of isolated words or decontextualized sentences, as in the study by Peterson & Barney (1952). Notable exceptions are two papers by Umeda (1975, 1977). The first presented a study of vowel durations in "10- to 20-min readings by three different speakers," comprising a total of about 45 minutes of speech; the second presented a study of consonant duration in one of those readings, comprising about 20 minutes of speech. Both of Umeda's studies were aimed at deriving parameters for text-to-speech synthesis. The source material was fluent reading of coherent text, rather than performance of isolated words or phrases. But again, only selected summary statistics were published, and in this case, the raw measurements as well as the original recordings have apparently been lost.

2.2. The Second Wave: Digital Methods

We can estimate the size of the Peterson and Barney data set due to a replication and extension by Hillenbrand et al. (1995), in which 139 speakers read 12 vowels each, for a total of 1,668 syllables measured. Including the carrier phrases (e.g., "Now say h_d again") and surrounding silence, the recorded audio for this study comprises nearly 31 minutes—but the measured syllables amounted to only 458 seconds, or approximately 7.5 minutes. In keeping with developing modern norms, Hillenbrand et al. published in digital form the complete audio data set used in their paper, as well as all of their measurements, perceptual results, and so forth. This publication—available as downloadable text files and .zip archives on Prof. Hillenbrand's website—was essentially free, and the digital form of the resulting data makes it easy for others to replicate and extend the results. This is a small example of the ways in which digital technology has transformed corpus-based research, in phonetics as well as in other areas of linguistics.

It is easier to publish digital audio recordings today than it was to publish paper books in 1950—Allen's 100 hours of Abaza would easily fit on a thumb drive costing less than five dollars, or could be downloaded in a few minutes from the internet. Millions of hours of real-world digital speech recordings are accumulating from sources such as audiobooks, podcasts, oral history collections, and political and legal archives. Freely available interactive speech analysis software produces better spectrograms (and more accurate measurements) than the expensive and slow analog sound spectrographs of 1950. And convenient software for organizing, searching, and analyzing large data sets turns months or years of human effort into minutes or hours of programming and interactive exploration.

2.2.1. Speech data sets for human language technology. In the late 1980s, the entry of the US Defense Advanced Research Projects Agency (DARPA) into the field of human language technology (HLT) promoted the creation and publication of larger acoustic–phonetic data sets. An early result was the Texas Instruments–MIT (TIMIT) corpus, created between 1987 and 1990 and documented by Garofolo et al. (1993). TIMIT consists of 6,300 sentences, 10 read by each of 630 speakers, totaling approximately 5.5 hours of audio, with time-aligned lexical and phonetic transcripts. The then-new availability of the CD-ROM delivery format allowed TIMIT to be published in a form that could be as easily and inexpensively distributed as a printed book, and as a result, thousands of copies were distributed. Google Scholar now lists more than 2,300 citations for Garofolo et al. 1993, and more than 18,000 publications that cite the TIMIT data set in other ways.

DARPA's HLT efforts were organized around a "common task" model, which aimed to produce reliable and reproducible progress through publication of three crucial items at the start of each project:

1. A detailed task definition and evaluation plan.
2. Automatic quantitative evaluation software, written and administered by the National Institute of Standards and Technologies.
3. A large body of shared training and "development test" data, with "evaluation test" data withheld for periodic public evaluation workshops.

This approach to R&D management was a massive success in promoting steady progress in text-based as well as speech-based technologies, and was widely copied in other domains and in other countries. One result was thousands of published speech data sets, and a flowering of organizations dedicated to this topic. The Linguistic Data Consortium (**http://ldc.upenn.edu**), formed in 1992 with seed money from DARPA to help create, curate, and distribute such data, has published 755 speech and language data sets and has distributed more than 140,000 copies of these

data sets to companies, universities, and government research laboratories in more than 80 countries (Cieri et al. 2018). The Bavarian Archive for Speech Signals (BAS; **https://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html**) was founded in 1995, "with the aim of making speech resources of contemporary spoken German as well as tools for the processing of digitized speech available to research and speech technology communities." BAS currently offers 37 speech data sets. The European Language Resources Association (ELRA; **http://www.elra.info**) was also founded in 1995 "to make Language Resources (LRs) for Human Language Technologies (HLT) available to the community at large." ELRA now offers 486 speech data sets. Many other technology-oriented digital speech data sets have been created and distributed by individual researchers or research organizations, various national or regional repositories, and several commercial enterprises such as Appen and Speech Ocean.

2.2.2. Use of human language technology data sets and analysis techniques in phonetics. Although the TIMIT data set described in the previous section was created to support engineering research in speech recognition, it was soon used for basic research in phonetics, as described in the abstract from Byrd (1992, p. 593):

> A set of phonetic studies based on analysis of the TIMIT speech database is presented. Using a database methodological approach, these studies detail new results in speaker-dependent variation due to sex and dialect region of the talker including effects on stop release frequency, speaking rate, vowel reduction, flapping, and the use of glottal stop. TIMIT was found to be fertile ground for gathering acoustic-phonetic knowledge having relevance to the phonetic classification and recognition goals for which TIMIT was designed as well as to the linguist attempting to describe regularity and variability in the pronunciation of read English speech.

Other early TIMIT-based papers covered vowel and consonant reduction (Manuel et al. 1992), the acoustic characteristics of stops (Byrd 1993), the relationship between gender and dialect (Byrd 1994), and word and segment duration (Keating et al. 1994).

Another early speech technology data set[1] was the HCRC Map Task corpus (Anderson et al. 1991), containing 128 task-oriented dialogues carried out among university students in Glasgow, Scotland. Each participant has their own individual copy of a schematic map, consisting of an outline and roughly a dozen labeled features. Most features are common to the two maps, but not all. One map has a route drawn in, while the other does not. The task is for the participant without the route to add it to their map, on the basis of discussion with the participant whose map has it. Again, although the main purpose of this data set was to support technological research and development, it was often used to investigate scientific issues as well. To this end, Boyle et al. (1994) explored the role of visibility and gaze, while Bard & Aylett (1999) explored the interaction among deaccenting, givenness, and syntactic role. The "map task" design has been widely copied, with similar collections produced in American English, Canadian English, Dutch, German, Japanese, Korean, Occitan, Polish, Portuguese, Swedish, Thai, and Vietnamese. These collections have also been used for scientific as well as technological research.

Switchboard (Godfrey et al. 1992) was a collection of about 2,400 American English telephone conversations among 543 speakers. The original purpose of this data set was to support research in speaker identification, but it was soon also used in speech recognition research—and, again,

[1]For links to the data sets described in this section, see the Related Resources.

in phonetics research, as in Greenberg et al.'s (1996) report on "Insights into Spoken Language Gleaned from Phonetic Transcription of the Switchboard Corpus."

The CALLHOME data sets, comprising 120 half-hour telephone conversations in each of six languages, were collected and published between 1994 and 1997. These collections were used to support multilingual conversational speech recognition research. Fox (2000) used the Spanish CALLHOME recordings to study the lenition of syllable-final /s/, where /s/ may be reduced in duration, partly or completely voiced, transformed by loss of oral frication into a kind of [h]-like sound, or entirely deleted. She used the published word-level alignments, but relied on two human annotators to code the treatment of each of 24,473 instances of syllable-final /s/.

An open-ended range of alternative measures are now accessible to anyone who can write simple computer programs. As a result, research reports today are based on data sets three or four orders of magnitude larger than those of only a few decades ago.

2.2.3. The development of (semi)automatic methods for alignment and measurement. An important by-product of speech technology research was a set of techniques for accurately aligning speech recordings with orthographic transcripts, in the process deriving aligned phonetic transcripts. Such transcripts enable detailed phonetic analysis of many hours of speech recordings with little or no human intervention, as they automatically tabulate measurements of standard phonetic variables like vowel quality or segment durations. Specifically, the input is digital text in standard orthographic form (whether a transcription or a source that was read) paired with a digital audio recording; the output consists of the start and end times of each word (or other orthographic unit), along with a similarly time-aligned sequence of phonetic segments.

The central idea in this research was the method of speech-to-text alignment used in training the acoustic models for hidden Markov model (HMM) speech recognition systems (Rabiner 1989). The basic idea of HMM systems is simple: We want to transform an audio recording automatically into ordinary text. The recording is an "observable" (indeed observed!) sequence of sounds, which corresponds to an unknown ("hidden") sequence of words. We aim to compute the hidden word sequence by jointly optimizing two quantities: the a priori probability of the word sequence (which is independent of the recording) and the probability of the sound sequence given the word sequence. These probabilities are estimated using statistical models whose parameters are estimated from a training set. Once we have those models, the Viterbi algorithm (Viterbi 1967) offers an efficient (linear-time) technique for determining the hidden word sequence most likely to correspond to a new acoustic sequence.

The a priori word-sequence probability is estimated using a statistical language model, typically an *n*-gram model, so called because it estimates the conditional probability of each word given the *n* previous words. This is a Markov model because it assumes that the sequential conditioned choices of words are statistically independent; and it is an HMM because the available observations are not the words themselves, but rather a sequence of sounds, which are viewed as a stochastic function of the hidden word sequence. This function, which estimates the conditional probability of the observed acoustics given a particular word sequence, is called the acoustic model. The acoustic model decomposes the word sequence into a sequence of phonetic segments called phones, corresponding roughly to the symbols in a dictionary pronunciation, and further divides each phone into a series of linked states, perhaps varied according to the phonetic context.

The language model and the acoustic models are trained separately. The language model is trained using the largest-available collection of relevant text (this aspect of the system is not discussed further in this review). The acoustic models are trained using a large collection of transcribed speech, typically between tens and thousands of hours.

If human phoneticians laboriously translated the training-set word sequence into a sequence of phonetic segments, and aligned each segment with the corresponding period of time in the audio recording, then estimating the parameters of the acoustic model would be easy. We would simply need to collect the audio for each phonetic segment across the whole training set, and estimate the corresponding set of probability distributions, using standard statistical techniques. But in the 1960s, Baum et al. (1970) at the Institute for Defense Analysis developed an iterative method, known as the Baum–Welch algorithm or the forward–backward algorithm, which in principle allows the parameters of an HMM to be estimated given only observable sequences.

In practical applications to speech technology, the training input includes texts paired with audio recordings. A pronunciation dictionary and so-called letter-to-sound rules are used to map the word sequence to a sequence of phones. The role of the forward–backward algorithm is to align the phone sequence and the audio recording, in the process estimating acoustic models for each of the phones. (As mentioned above, such systems usually make use of sequential substates for a much larger set of phones in context, for example, triphones, each of which consists of a phonetic segment preceded and followed by other specific segments. Much statistical complexity ensues, but the principle is unchanged.)

Once a set of acoustic models have been trained, we have a new practical opportunity. Given pairs of texts and corresponding audio recordings, we can treat the known texts as a particularly rigid sort of language model, and the regular HMM decoding techniques will produce an optimal alignment of the audio with those texts (and their inferred phone sequences). We can allow the language model to consider alternative word pronunciations, alternative dialect forms, and optional between-word silences, and the same techniques will determine the most probable transcription as well as the most probable alignment.

The first use of these techniques on a significant scale was in the semiautomatic lexical and phonetic alignment of the TIMIT corpus in the late 1980s. In that case, the automatically created alignments and hypothesized phone sequences (derived from dictionary pronunciations) were corrected by human annotators. Since then, software and models for forced alignment have become widely available and have been extensively used.

Since 1989, Steve Young and Phil Woodland of the Department of Engineering at the University of Cambridge have produced and distributed a series of versions of the Hidden Markov Model Toolkit (HTK), which is a set of C library modules and tools meant for speech recognition research. The accompanying HTKBook not only documents the use of the programs but also explains the underlying theory in detail (**http://htk.eng.cam.ac.uk/docs/docs.shtml**). Several wrappers adapting HTK for forced alignment have been developed and made available as open-source software—an early example was the Penn Phonetics Lab Forced Aligner. Starting in 1997, Lee et al. (2001) in Japan have developed and distributed Julius, an open-source HMM system which forms the foundation of the SPPAS phonetic tool kit developed by Bigi (2015) in France. Martin's (2004) Winpitch program integrates forced alignment into a system intended for prosodic analysis, especially of Romance languages (see also Cresti et al. 2004). A more recent speech recognition tool kit is Kaldi (Povey et al. 2011), which has also been adapted specifically for forced alignment in several other projects, notably as the Montreal Forced Aligner (Gorman et al. 2011).

The alignments produced by such systems can be quite accurate. Stolcke et al. (2014) found that the phone boundaries in their automatically derived alignments for the TIMIT acoustic–phonetic data set corresponded within 20 milliseconds to the boundaries placed by human phoneticians 96.8% of the time. But this level of accuracy depends on accurate "pronunciation modeling," that is, determining how the transcribed word sequence was actually pronounced in the examples being analyzed. A remaining challenge for the field of corpus phonetics is the development of

effective and accurate pronunciation-modeling techniques that can be applied across languages and varieties.

Lacking reliable general methods for this type of pronunciation modeling, researchers have several options. One is to use forced-alignment methods to locate examples of interest, and then classify and/or measure them by hand. With appropriate software support, this process can be quite efficient, requiring only a few seconds of annotator time per token, so that four or five thousand tokens can be processed in a day's work. Fox (2000) used a version of this method, described above.

Another approach is to devise special-purpose classification algorithms for particular cases. Examples include an algorithm for voice onset time (Sonderegger & Keshet 2010), one for the clear- versus dark-/l/ continuum (Yuan & Liberman 2009, 2011a; described more fully in Section 2.4), one for vowel nasalization (Yuan & Liberman 2011b), one for *g*-dropping (Yuan & Liberman 2011c), and one for /s/-lenition (Ryant & Liberman 2016b).

Tools such as EXMARaLDA (Schmidt 2004; **http://exmaralda.org/de/**), LaBB-CAT (Fromont & Hay 2012; **https://labbcat.canterbury.ac.nz/system/**), and EMU-SDMS (Winkelmann et al. 2017; **https://ips-lmu.github.io/EMU.html**) aim to various extents to integrate alignment, speech signal processing, database management, search, and statistical analysis for spoken corpora. As explained below, we can expect that future systems of this type will be more widely used, offering convenient access to distributed data sets through a consistent interface.

2.3. Scientific and Humanistic Speech Data Set Projects

In addition to the proliferation of data set publications for speech technology purposes, the period since the mid-1980s has seen the creation of many scientific and humanistic projects focused in whole or in part on collections of speech and language data for research purposes. A survey of all of these projects would easily fill another review article, so this section discusses only a representative sample.

In 1984, Brian MacWhinney and Catherine Snow established CHILDES (MacWhinney 1996) as a repository for child language acquisition data, and during the following decades, the research community engaged with child language acquisition came to accept sharing of transcripts (and sometimes recordings) as the norm in their field. More recently, TalkBank (**http://talkbank.org**) was developed as an online location for speech and language data related to second-language acquisition, conversation analysis, classroom discourse, and several clinical categories (MacWhinney 2001).

William Kretzschmar, at the University of Georgia, has collected paper documentation and tape recordings from 10 large-scale dialect survey projects carried out in various regions of the United States between the 1930s and the present (Kretzschmar 2003). Some but not all of this material has been digitized and is available on the Linguistic Atlas Projects website (**http://us.english.uga.edu/cgi-bin/lapsite.fcgi**).

PennSound, founded at the University of Pennsylvania by Charles Bernstein and Al Filreis, is an ongoing project "committed to producing new audio recordings and preserving existing audio archives" for readings of literary works and interviews with authors (Bernstein 2003). The online archive now includes nearly 6,000 readings by more than 700 authors.

The Sociolinguistic Archive and Analysis Project (SLAAP) was established in 2007 at North Carolina State University (Kendall 2007). According to the project website (**https://slaap.chass.ncsu.edu/**), as of January 2018 the SLAAP archive contained more than 4,450 interviews comprising 3,850 hours of audio. A total of 190 hours of this material has been orthographically transcribed, amounting to about 1.85 million words, "accurately time-stamped and linked to the audio from a variety of languages (predominately American dialects in North Carolina and the southeastern United States)."

The Common Language Resources and Technology Infrastructure (CLARIN) was established in 2012 to "to create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences" (Hinrichs & Krauwer 2014). CLARIN is primarily a European project, with more than 20 local centers, such as the BAS (described in Section 2.2.1, above), Språkbanken in Sweden (**https://spraakbanken.gu.se**) and the Instituut voor de Nederlandse Taal in the Netherlands (**http://ivdnt.org**). The CLARIN Virtual Language Observatory (**https://vlo.clarin.eu/**) provides links to numerous local repositories and currently indexes more than 160,000 spoken data sets. Their size, availability, and generality of interest vary widely, but the count itself is impressive.

2.4. "Found Data"

Increasingly large amounts of digitally recorded spoken material are becoming available as a normal part of modern social life. Examples include broadcast sources (news and interview programs, talk shows, "reality" shows, etc.) as well as podcasts, audiobooks, political speeches and debates, and court proceedings. In recent years, phoneticians have begun to use these collections of "found data" as a basis for their research.

Yuan & Liberman (2009) used aligned recordings from oral arguments presented during the 2001 term of the US Supreme Court to study a measure of variation in the pronunciation of /l/, a difference traditionally described as clear (in syllable-initial prestress position) versus dark (in syllable- and word-final position). Sproat & Fujimura (1993), in a study using both articulatory and acoustic evidence, argued that the clear/dark distinction should not be treated as a categorical difference, but rather as a gradient correlate of a continuum of syllabic affinity. Yuan and Liberman investigated the same question, using likelihood ratios from the fit of "clearly clear" and "clearly dark" instances, partly confirming and partly contradicting Sproat and Fujimura's conclusions. A crucial difference was that Yuan & Liberman (2009) examined 21,706 tokens of /l/, nearly 100 times as many as did Sproat & Fujimura (1993), which allowed new sets of questions to be addressed. In addition, the 2009 study looked at data from naturally occurring spontaneous speech, removing the possibility of contextual effects from subjects reading lists of sentences in which a single dimension is systematically varied. Note that the archive of US Supreme Court oral arguments now includes all audio recorded in the Court since October 1955, amounting to more than 14,000 hours of audio and more than 66 million words (**http://oyez.org**).

As another example, the LibriVox archive of public-domain audiobooks (**https://librivox.org/**) now includes more than 60,000 hours of English-language recordings, with links to the associated texts, and a smaller but still considerable number of recordings in many other languages. Panayotov et al. (2015) selected a small subset of this material and published it as the LibriSpeech corpus, comprising 5,832 chapters read by 2,484 speakers, with a total duration of more than 1,570 hours. The LibriSpeech collection has mainly been used for speech technology investigations, but it is starting to be used in phonetics research with a more scientific orientation (e.g., Ryant & Liberman 2016a, Chodroff & Wilson 2017).

3. THE FUTURE: CORPUS PHONETICS IN 2030

Since the underlying technology and the associated science of corpus phonetics are changing so rapidly, it would be fun but foolish to try to project more than a dozen or so years into the future. But that span of time will already be enough to see an important range of quantitative and qualitative changes in this general area. The following subsections describe five such changes in turn.

3.1. More Acoustic Data

We can expect a continued exponential increase in available data for acoustic–phonetic research. Some of these data will come from the ongoing digitization of dusty stacks of analog tapes, as in the cases of the University of Georgia Linguistic Atlas Projects mentioned above; the Labov Archive of about 9,000 sociolinguistic interviews now being processed at the University of Pennsylvania; and the WFMT archive of Studs Terkel radio interviews (**https://studsterkel.wfmt.com/**), 1,200 of which are now available online. But most of the data will come from the natural accumulation of new broadcasts, podcasts, audiobooks, oral histories, and so forth, as well as from the spread of the Open Data movement through speech-related research disciplines. There are already tens of thousands of hours of transcribed and aligned English-language audio available to the broader research community—large companies like Google and Amazon have internal access to a million hours or more, and a substantial fraction of these recordings will undoubtedly become more generally available over the next dozen years, facilitating detailed corpus-based studies of allophonic and prosodic variation across style, register, age, rhetorical and information structure, and so on, as well as dialectology on an unprecedentedly large scale.

We can expect other major world languages to reach a comparable level over the next decade. Underdocumented languages with large speaker populations will routinely come to have thousands of hours of available acoustic–phonetic data sets, and techniques are being developed to create hundreds of hours of acoustic data for small and endangered languages (e.g., Bird et al. 2014, Blachon et al. 2016). This proliferation of acoustic data across languages will permit serious corpus-based comparative studies, not only of segmental and prosodic phonetics but also of cultural differences in the phonetic correlates of self-presentation and communicative interaction (e.g., Liberman 2007, Yuan et al. 2007).

3.2. A Larger Number of More Sophisticated Users

We can expect that the knowledge and skills necessary for research in corpus phonetics will spread more widely among linguists in various subdisciplines, as well as among researchers in fields such as psychology, anthropology, sociology, medicine, literary studies, law, and political science. In part, this process will reflect the general trend toward more computational sophistication in the research community at large, and in part it will result from better tools and better documentation. And a virtuous circle will result, in which a larger user community creates larger bodies of interesting data and a larger collection of interesting results, thereby recruiting even more participants.

3.3. Articulatory, Physiological, and Perceptual Data

Most articulatory, physiological, and perceptual data sets are relatively small and are held as the private property of their creators, but both of these things are likely to change, since much of the necessary instrumentation is becoming less expensive and easier to use, and the Open Data movement is starting to change attitudes in these fields as well. It is too early to document much of a trend in this area, but we can predict that things will look quite different in 2030. We can thus expect an increase in the availability of various articulatory, physiological, and perceptual data sets.

3.4. Ease of Access: Standards, Tools, and Distributed Storage

As mentioned in Section 2.2, new tools such as EMU-SDMS combine annotation, database search, and statistical analysis in a web-based application. In principle, this means that users will no longer

need to have a local copy of every data set they want to analyze, or to process each data set to fit their suite of analysis programs. Instead, they will be able to attach copies "in the cloud" of any data sets they have credentials for and analyze them with standard functions, or use a standard access API to run their own analysis programs. The fact that cloud services like AWS and Azure offer storage in compliance with the Health Insurance Portability and Accountability Act (HIPAA) and the Family Educational Rights and Privacy Act (FERPA) will weaken some of the arguments against sharing of publicly funded data sets. The infrastructure for assigning and using access credentials does not yet exist for routine use in this scenario, but it surely will become available over the next dozen years, giving rise to the ability to attach and explore data sets as easily as we now explore news stories or stream video content.

The growing use of executable "notebooks" like knitr (Xie 2015) and Jupyter (Kluyver et al. 2016) will combine in a fruitful way with this distributed data ecosystem, giving readers (and authors) the ability to replicate the steps involved in a published analysis by simply clicking on an icon. We can expect that by 2030, all reputable journals will require such executable notebook instances to be supplied for every published paper, in corpus phonetics as well as in all other areas of computational analysis of the natural world.

3.5. More Complete Automation

Finally, a series of developments will in effect create a "robot phonetician" or, perhaps, a robot phonetician's assistant. As discussed above, researchers have already taken the first steps in that direction with the automatic analysis of Spanish /s/-lenition (Fox 2006), vowel formant measurements (Evanini et al. 2009), /l/-allophony (Yuan & Liberman 2009), voice onset time (Sonderegger & Keshet 2010), *g*-dropping (Yuan & Liberman 2011b), and so on. But each of these investigations required a dedicated algorithm, implemented and tested on a particular task (and often on a particular data set).

Well before 2030, we can expect to see a tool that will automatically evaluate, classify, and measure all aspects of the phonetic interpretation of arbitrary English speech, given an accurate transcript. After some additional effort, this tool will work across dialects, styles, and recording conditions, with roughly the quality of a phonetically well-trained research assistant. At about the same time, we will see specialized systems that can perform the analogous task for other major languages. Somewhat later, there may be a generalized system that can work across languages, again with the same sort of quality expected from the analysis of a well-trained phonetician encountering material in a new language along with dictionary-style pronunciation information for the words. And at some point in this process, speech recognition will work well enough that such systems can produce useful results without a transcript, at least for languages for which adequate automatic speech recognition systems have been trained.

4. CONCLUSION

We have already seen the leading edge of a cultural shift in scientific studies of speech and language toward empirical analysis of very large collections of real-world text and speech. The forces driving this shift are the same technological forces that are influencing other areas of science: the availability of large quantities of digital data and the development of algorithms and computer tools that make it increasingly easy and cheap to process increasingly large data sets. The change has been especially strong in the case of phonetics, since it involves quantitative modeling of highly variable multidimensional patterns. Researchers began doing corpus-based phonetics in a limited way a century ago, and as electronic and digital technologies have advanced, phoneticians have

been among the early adopters of each wave of new methods. We can confidently predict that these trends will accelerate over the next few years, as new algorithms, new interfaces, and new cultural norms make it less costly, faster, and easier to explore new descriptive domains; to test new empirical hypotheses; and to contest, validate, or extend the results of others.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Allen WS. 1956. Structure and system in the Abaza verbal complex. *Trans. Philol. Soc.* 55:127–76

Anderson AH, Bader M, Bard EG, Boyle EA, Doherty G, et al. 1991. The HCRC map task corpus. *Lang. Speech* 34:351–66

Bard EG, Aylett MP. 1999. The dissociation of deaccenting, givenness, and syntactic role in spontaneous speech. In *Proceedings of the 14th International Congress of Phonetic Sciences* (*ICPhS-14*), pp. 1753–56. London: Int. Phon. Assoc.

Baum LE, Petrie T, Soules G, Weiss N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* 41:164–71

Bernstein C. 2003. *PennSound Manifesto*, Univ. Pa., Philadelphia. **http://writing.upenn.edu/pennsound/manifesto.php**

Bigi B. 2015. SPPAS-multi-lingual approaches to the automatic annotation of speech. *Phonetician* 111/112:54–69

Bird S, Hanke FR, Adams O, Lee H. 2014. Aikuma: a mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 1–5. Stroudsburg, PA: Assoc. Comput. Linguist.

Blachon D, Gauthier E, Besacier L, Kouarata GN, Adda-Decker M, Rialland A. 2016. Parallel speech collection for under-resourced language studies using the Lig-Aikuma mobile device app. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association* (*INTERSPEECH2016*), pp. 61–66. Red Hook, NY: Curran

Boyle EA, Anderson AH, Newlands A. 1994. The effects of visibility on dialogue and performance in a cooperative problem solving task. *Lang. Speech* 37:1–20

Byrd D. 1992. Preliminary results on speaker-dependent variation in the TIMIT database. *J. Acoust. Soc. Am.* 92:593–96

Byrd D. 1993. 54,000 American stops. *UCLA Work. Pap. Phon.* 83:97–116

Byrd D. 1994. Relations of sex and dialect to reduction. *Speech Commun.* 15:39–54

Chiba T, Kajiyama M. 1941. *The Vowel: Its Nature and Structure*. Tokyo: Kaiseikan

Chodroff E, Wilson C. 2017. Structure in talker-specific phonetic realization: covariation of stop consonant VOT in American English. *J. Phon.* 61:30–47

Cieri C, Liberman M, Strassel S, DiPersio D, Wright J, et al. 2018. From 'solved problems' to new challenges: a report on LDC activities. In *Proceedings of the 11th Conference in International Language Resources and Evaluation* (*LREC18*), pp. 3265–69. Paris: Eur. Lang. Resour. Assoc.

Cordereix P. 2014. Ferdinand Brunot et *Les Archives de la parole*: le phonographe, la mort, la mémoire. *Revue BNF* 48:5–11

Cresti E, do Nascimento FB, Moreno-Sandoval A, Veronis J, Martin P, Choukri K. 2004. The C-ORAL-ROM CORPUS: a multilingual resource of spontaneous speech for romance languages. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (*LREC04*), pp. 575–78. Paris: Eur. Lang. Resour. Assoc.

Ellis AJ. 1874. On the physical constituents of accent and emphasis. *Trans. Philol. Soc.* 15:113–64

Evanini K, Isard S, Liberman M. 2009. Automatic formant extraction for sociolinguistic analysis of large corpora. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association* (*INTERSPEECH2009*), pp. 1655–58. Baixas, Fr.: Int. Speech Commun. Assoc.

Fant G. 1956. On the predictability of formant levels and spectrum envelopes from formant frequencies. In *For Roman Jakobson: Essays on the Occasion of His Sixtieth Birthday*, ed. M Halle, HG Lunt, CH Van Schooneveld, pp. 109–20. The Hague: Mouton

Fox MA. 2000. Syllable-final /s/ lenition in the LDC's CallHome Spanish Corpus. In *Proceedings of the 6th International Conference on Spoken Language Processing* (*ICSLP2000*), pp. 556–59. Beijing: China Mil. Friendsh. Publ.

Fox MA. 2006. *Usage-based effects in Latin American Spanish syllable-final /s/ lenition*. PhD thesis, Univ. Pa., Philadelphia

Fromont R, Hay J. 2012. LaBB-CAT: an annotation store. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pp. 113–17. Stroudsburg, PA: Assoc. Comput. Linguist.

Garofolo J, Lamel L, Fisher W, Fiscus J, Pallett D, Dahlgren N. 1993. *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*. CD-ROM, Natl. Inst. Stand. Technol. rep. 4930, Washington, DC

Godfrey JJ, Holliman EC, McDaniel J. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), pp. 517–20. Piscataway, NJ: IEEE

Gorman K, Howell J, Wagner M. 2011. Prosodylab-aligner: a tool for forced alignment of laboratory speech. *Can. Acoust.* 39:192–93

Greenberg S, Hollenback J, Ellis D. 1996. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In *Proceedings of the 4th International Conference on Spoken Language Processing* (*ICSLP96*), pp. 24–27. Newark: Speech Res. Lab., Univ. Del.

Hillenbrand J, Getty LA, Clark MJ, Wheeler K. 1995. Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97:3099–111

Hinrichs E, Krauwer S. 2014. The CLARIN research infrastructure: resources and tools for eHumanities scholars. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (*LREC14*), pp. 1525–31. Paris: Eur. Lang. Resour. Assoc.

Jones D. 1909. *Intonation Curves: A Collection of Phonetic Texts, in Which Intonation Is Marked Throughout by Means of Curved Lines on a Musical Stave*. Berlin: Teubner

Keating PA, Byrd D, Flemming E, Todaka Y. 1994. Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Commun.* 14:131–42

Kendall T. 2007. Enhancing sociolinguistic data collections: the North Carolina Sociolinguistic Archive and Analysis Project. *Univ. Pa. Work. Pap. Linguist.* 13:15–26

Kendall T. 2013. *Speech Rate, Pause and Sociolinguistic Variation: Studies in Corpus Sociophonetics*. Berlin: Springer

Kitchin R. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Thousand Oaks, CA: Sage

Kirkham S, Moore E. 2016. Constructing social meaning in political discourse: phonetic variation and verb processes in Ed Miliband's speeches. *Lang. Soc.* 45:87–111

Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, et al. 2016. Jupyter Notebooks—a publishing format for reproducible computational workflows. In *Proceedings of the 20th International Conference on Electronic Publishing*, pp. 87–90. Amsterdam: IOS

Koenig W, Dunn HK, Lacy LY. 1946. The sound spectrograph. *J. Acoust. Soc. Am.* 18:19–49

Kretzschmar WA. 1993. *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*. Chicago: Univ. Chicago Press

Kretzschmar WA. 2003. Linguistic atlases of the United States and Canada. *Am. Speech* 88:25–48

Labov W. 1963. The social motivation of a sound change. *Word* 19:273–309

Labov W, Rosenfelder I, Fruehwald J. 2013. One hundred years of sound change in Philadelphia: linear incrementation, reversal, and reanalysis. *Language* 89:30–65

Lee A, Kawahara T, Shikano K. 2001. Julius—an open source real-time large vocabulary recognition engine. In *Proceedings of the 2nd International Conference on Speech Communication and Technology* (*INTERSPEECH2001*), pp. 1691–94. Baixas, Fr.: Int. Speech Commun. Assoc.

Leslie J, Snyder R. 2010. *History of the early days of Ampex Corporation*. Paper, Audio Eng. Soc. (AES) Hist. Comm., AES, New York

Liberman M. 2007. Nationality, gender, and pitch. *Language Log Blog*, Nov. 12. **http://itre.cis.upenn.edu/~myl/languagelog/archives/005104.html**

MacArthur MJ. 2016. Monotony, the churches of poetry reading, and sound studies. *PMLA* 131:38–63

MacWhinney B. 1996. The CHILDES system. *Am. J. Speech Lang. Pathol.* 5:5–14

MacWhinney B. 2001. From CHILDES to TalkBank. In *Research on Child Language Acquisition*, ed. M Almgren, A Barreña, M Ezeizaberrena, I Idiazabal, B MacWhinney, pp. 17–34. Somerville, MA: Cascadilla

Manuel SY, Shattuck-Hufnagel S, Huffman MK, Stevens KN, Carlson R, Hunnicutt S. 1992. Studies of vowel and consonant reduction. In *Proceedings of the 2nd International Conference on Spoken Language Processing* (*ICSLP92*), pp. 943–46. Edmonton: Univ. Alberta

Martin P. 2004. Winpitch corpus, a text to speech alignment tool for multimodal corpora. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (*LREC04*), pp. 537–40. Paris: Eur. Lang. Resour. Assoc.

Meyer AS, Huettig F, Levelt WJ. 2016. Same, different, or closely related: What is the relationship between language production and comprehension? *J. Mem. Lang.* 89:1–7

Panayotov V, Chen G, Povey D, Khudanpur S. 2015. Librispeech: an ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), pp. 5206–10. Piscataway, NJ: IEEE

Peterson GE, Barney HL. 1952. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24:175–84

Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, et al. 2011. The Kaldi speech recognition toolkit. In *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 158–63. Piscataway, NJ: IEEE

Rabiner LR. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77:257–86

Ryant N, Liberman M. 2016a. Automatic analysis of phonetic speech style dimensions. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association* (*INTERSPEECH2016*), pp. 77–81. Red Hook, NY: Curran

Ryant N, Liberman M. 2016b. Large-scale analysis of Spanish /s/-lenition using audiobooks. In *Proceedings of the 22nd International Congress on Acoustics*, pap. ICA2016-721. Buenos Aires: MCI

Schmidt T. 2004. Transcribing and annotating spoken language with EXMARaLDA. In *Proceedings of the LREC Workshop on XML-Based Richly Annotated Corpora*, pp. 69–74. Paris: Eur. Lang. Resour. Assoc.

Sonderegger M, Keshet J. 2010. Automatic discriminative measurement of voice onset time. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association* (*INTERSPEECH2010*), pp. 2242–45. Baixas, Fr.: Int. Speech Commun. Assoc.

Sproat R, Fujimura O. 1993. Allophonic variation in English /l/ and its implications for phonetic implementation. *J. Phon.* 21:291–311

Stolcke A, Ryant N, Mitra V, Yuan J, Wang W, Liberman M. 2014. Highly accurate phonetic segmentation using boundary correction models and system fusion. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), pp. 5552–56. Piscataway, NJ: IEEE

Umeda N. 1975. Vowel duration in American English. *J. Acoust. Soc. Am.* 58:434–45

Umeda N. 1977. Consonant duration in American English. *J. Acoust. Soc. Am.* 61:846–58

Viterbi A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* 13:260–69

Watrous RL. 1991. Current status of Peterson–Barney vowel formant data. *J. Acoust. Soc. Am.* 89:2459–60

Winkelmann R, Harrington J, Jänsch K. 2017. EMU-SDMS: advanced speech database management and analysis in R. *Comput. Speech Lang.* 45:392–410

Xie Y. 2015. *Dynamic Documents with R and knitr*. Boca Raton, FL: CRC

Yuan J, Liberman M. 2009. Investigating /l/ variation in English through forced alignment. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association* (*INTERSPEECH2009*), pp. 2215–18. Baixas, Fr.: Int. Speech Commun. Assoc.

Yuan J, Liberman M. 2011a. /l/ variation in American English: a corpus approach. *J. Speech Sci.* 1:35–46

Yuan J, Liberman M. 2011b. Automatic detection of "g-dropping" in American English using forced alignment. In *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 490–93. Piscataway, NJ: IEEE

Yuan J, Liberman M. 2011c. Automatic measurement and comparison of vowel nasalization across languages. In *Proceedings of the 17th International Congress of Phonetic Sciences* (*ICPhS XVII*), pp. 2244–47, London: Int. Phon. Assoc.

Yuan J, Liberman M, Cieri C. 2007. Towards an integrated understanding of speech overlaps in conversation. In *Proceedings of the 16th International Congress of Phonetic Sciences* (*ICPhS XVI*), pp. 1337–40. London: Int. Phon. Assoc.

RELATED RESOURCES

1. CALLHOME Spanish Transcripts. **https://catalog.ldc.upenn.edu/LDC96T17**
2. HCRC Map Task Corpus. **http://groups.inf.ed.ac.uk/maptask/**
3. Switchboard-1 Telephone Speech Corpus. **https://catalog.ldc.upenn.edu/ldc97s62**

Annual Review
of Linguistics

Volume 5, 2019

Contents

Errata

An online log of corrections to *Annual Review of Linguistics* articles may be found at http://www.annualreviews.org/errata/linguistics