# Managing data for integrated speech corpus analysis in *SPeech Across Dialects of English* (SPADE)

Morgan Sonderegger, Jane Stuart-Smith, Michael McAuliffe, Rachel Macdonald, Tyler Kendall

October 2019 version of chapter to appear in *Open Handbook of Linguistic Data Management* (MIT Press)

## 1. Preliminaries: Large-scale speech corpus analysis

This Data Management Use Case discusses the SPeech Across Dialects of English (SPADE) project (for details, see www.spade.arts.gla.ac.uk). SPADE was devised to carry out large-scale integrated speech corpus analysis across a subset of Englishes. In so doing, the project aims to facilitate large-scale integrated speech corpus analysis for the speech and linguistics research communities in two ways. First, the project will generate large, publically-available, derived datasets of acoustic measures for English speech sounds. Second, it will create freely accessible software for future use by other researchers for analysing their own datasets: the Integrated Speech Corpus Analysis (ISCAN) system. The intended audience for this chapter thus includes readers wishing to use the derived datasets of acoustic measures in their own work and seeking background on the SPADE project and how the ISCAN software works; and readers who would like to carry out their own large-scale integrated corpus analysis projects (as part of a research team), and would like to know about a previous effort.

The vision behind SPADE is to enable less and more experienced users to carry out large-scale automatic search and extraction of the same information about speech from numerous spoken corpora. The user should be able to carry out this analysis whether the corpus is public or private, and independent of the corpus format, structure, complexity, and the dialect(s) it represents. As of October 2019, the project is two years in, and has laid the groundwork by developing ISCAN (Section 3). We are refining and testing this software for subsets of a single language, English, as represented by some 40 existing public and private spoken datasets from the Old World (British Isles) and New World (North America) across an effective time span of over 100 years. SPADE's research remit is to use ISCAN to investigate how segmental features of English, in particular vowels, sibilants, stops, and liquids, have changed over time and space.

SPADE was motivated by the desire to marry the availability of spoken language corpora with increasingly advanced speech processing tools, and to make feasible the sharing of existing speech datasets through robust automated speech analysis. There are now vast resources of digital collections of transcribed speech, from many different languages, gathered for many different purposes: from oral histories to sociolinguistic interviews, large datasets for training speech recognition systems, legal interactions, and political debates. The benefits of being able to share diverse speech corpora for the high-quality automated acoustic analysis of *spoken* as well as written language have implications within and beyond speech and linguistic research, including technological, forensic, and clinical approaches (cf. Liberman 2019). This is especially so if such analyses are standardized, replicable, and *ethically non-invasive*, i.e. can produce

anonymised acoustic measures or linguistic information (e.g. vowel formant measures or word frequencies), without the need for manual inspection or listening to speech from ethically-restricted spoken corpora. However, notwithstanding cost and privacy, there are numerous barriers to sharing speech corpora, including the nature of the speech datasets themselves in terms of size, complexity, and diversity of storage formats.

The availability of digital speech datasets is matched by the availability of increasingly complex speech processing tools. Automatic Speech Recognition-based tools for *forced alignment* automatically segment and label speech recordings which have written transcriptions, resulting in word- and sound-level boundaries. These tools have become increasingly widely used over the past decade (e.g. *FAVE*, *Montreal Forced Aligner*, *LaBB-CAT*, *MAUS*: Rosenfelder et al. 2015; McAuliffe et al. 2017a; Fromont and Hay 2012; Kisler et al. 2012), resulting in greatly-reduced search time for 'force-aligned' datasets. Machine learning-based software packages now allow for *automatic measurement* of some measures widely used in phonetic research (e.g. FAVE for vowel formants, *AutoVOT* for Voice Onset Time: Rosenfelder et al. 2015; Keshet et al. 2014). However, again, barriers prevent these tools from being widely used. Tools are generally specialized to particular dataset formats, and often require significant technical skill. For example, forced aligners require integration with electronic dictionaries which specify possible pronunciations of words, whilst measurement tools require command-line usage and some scripting in several programming languages (Python, R, Praat). Widely available speech analysis software such as *Praat* (Boersma and Weenik 2016) and *EMU-SDMS* (Winkelmann et al. 2017) also allows users to write their own programs (scripts) for the semi- and fully automatic measurement of some simple acoustic measures reflecting pitch, loudness, and noise components of speech (e.g. f0, amplitude, spectra), based on pre-implemented signal processing algorithms. But equivalent scripts are often written over and over again by different researchers, which has the methodological implication – reaching into theoretical inferences – that different acoustic analyses of the 'same' aspect of speech sounds are not actually the same.

The ISCAN system developed within SPADE forms part of a general movement towards development of different speech database management systems (e.g. EMU-SDMS, LaBB-CAT, *Phon*, *SLAAP*: Rose et al. 2006; Kendall 2007). These differ in their goals and functionality, depending on intended use cases. The ISCAN system for SPADE is specialized for linking and analysing multiple speech corpora, with flexibility for different use cases that do not assume users can necessarily access raw data.

ISCAN assumes that data annotation has been completed and performs data processing which can be carried out automatically, and which does not necessarily require manual/visual access to raw speech/text data; though an additional 'inspection interface' does permit access to raw audio provided the user has the appropriate permissions. ISCAN requires minimally a collection of sound files, with accompanying word and segment/phone-level time-stamped labelling (e.g. from forced alignment). Our approach to automated speech analysis assumes an abstraction away from the original speech dataset format, whereby raw audio + text datasets are imported and enriched with a large range of acoustic measures, resulting in anonymized databases of acoustic measures with additional linguistic information. These datasets can then be queried, and the results exported, resulting in 'derived datasets' (in CSV/spreadsheet format), for subsequent analysis (see Figure 1). Depending on requisite user permissions, additional functionality is also available for token by token inspection of raw audio. Our workflow provides standardized, customizable linguistic and acoustic measures across speech datasets, which in turn

will make reproducing and replicating investigation of speech much easier (Chapter 2, this volume).

The structure of our chapter is as follows. In Section 2, we discuss SPADE's approach to *data sharing*, in terms of data collection and working with our Data Guardians; the *citation and acknowledgement* of datasets – both primary audio corpora, and secondary, derived datasets of speech measures produced by ISCAN; and *data archiving*. In Section 3 we discuss the technical workflow for ISCAN. This covers the core aspects of *data processing and storage* for SPADE, which result in standardized databases for each speech corpus, and *access* and *search* of these databases, from which users generate the derived datasets. To keep the discussion concrete we exemplify using a SPADE project study, Stuart-Smith et al. 2019, on English /s/-retraction, whereby /s/ in /str/ clusters (e.g. *street*) sounds more like /sh/. The research question is: to what extent is /s/ acoustically 'retracted' relative to /ʃ/, as a function of onset structure (e.g. /sp st sk spr skr str/)? The data considered were from speakers of nine English dialects, from six spontaneous speech corpora (Section 2).

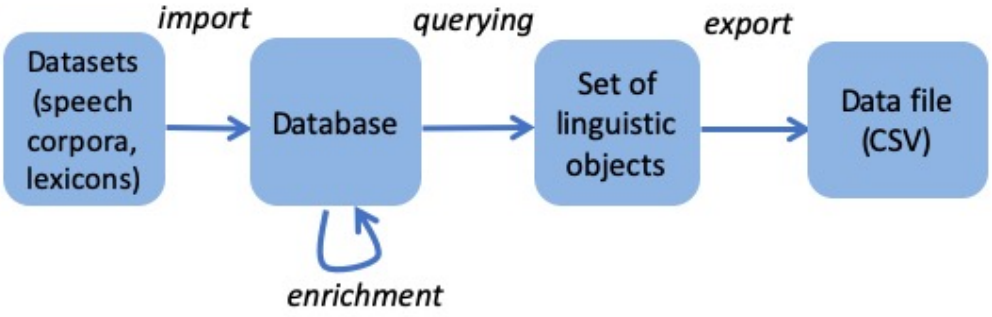


**Figure 1. Data processing workflow for SPADE. Details are given in Section 3.**

# 2. Data sharing in SPADE

## *2.1. Sampling*

Dataset sampling for SPADE was constrained by theoretical factors of dialect coverage by *space* and *time*, and in particular, the aim of capturing the relationship between British English dialects from the Old World with those continued in the New World, in North American dialects of Canada and the United States. Spatial dialect coverage was determined by design and availability (Figure 2). For the British Isles, we have standard and vernacular recordings from Ireland, Wales, Scotland, and England, and within England, Northern and Southern Englishes, as well as urban (e.g. London, Liverpool, Newcastle) and rural varieties (e.g. Devon, North-East Scotland), ethnic varieties (e.g. Glasgow Asian, Bradford Panjabi English), and varieties with known historical links with North American English, such as West Country and East Anglian English. For North America we aimed to cover the main recognized dialect regions following e.g. Thomas 2001. In terms of time, the speech datasets mainly date from the 1960s to the present. Many datasets contain either an apparent- or a real-time dimension, i.e. time depth is represented

through recordings made at different times, or through containing speakers recorded at the same time point, but of different ages.

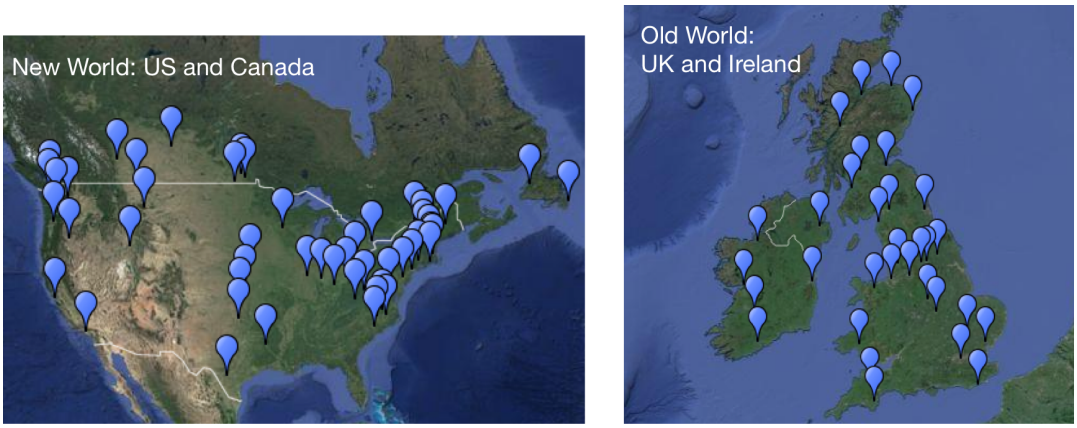


**Figure 2. Approximate locations of the intended SPADE dialect coverage.**

The speech datasets also vary by *speech style*. Most are largely of spontaneous speech, though some contain read speech (e.g. the International Corpus of English-Canada (ICE-Canada): Newman and Columbus 2010) or are entirely read speech (e.g. the Intonational Variation in English corpus: Grabe et al. 2001). The spontaneous speech recordings range from interviews of different kinds (oral history, sociolinguistic, broadcast, police-suspect roleplays), with differing numbers of speakers, to casual conversations with no interviewer present.

## *2.2. Data collection*

The SPADE speech datasets fall into two main types, *public* and *private*. The public datasets include: the Audio British National Corpus (AudioBNC: Coleman et al. 2012), the Buckeye Corpus (Pitt et al. 2007), the Dynamic Variability in Speech (DyViS) forensic corpus (Nolan et al. 2009), the Santa Barbara Corpus (Du Bois et al. 2004), and the Scottish Corpus of Texts and Speech (SCOTS: Anderson et al. 2007). These corpora are either freely accessible or available for sharing via a fee. Some of the large public corpora include recordings from many different dialects (e.g. AudioBNC, Santa Barbara, SCOTS), while others were collected from a single dialect area (e.g. Buckeye, DyViS). Hence speech corpus does not necessarily equate to English dialect.

The private datasets are largely those which have been collected for a specific purpose, often sociolinguistic or phonetic, and from one or a few dialects. They can also be internally complex. For example, the Sounds of the City corpus (SoTC: Stuart-Smith et al. 2017), is a real-time corpus of Glaswegian English recorded from the 1970s to the 2000s, which itself consists of: three sociolinguistic corpora, several sets of oral history interviews, broadcast recordings, and some other recordings. The Raleigh Corpus (Dodsworth and Kohn 2012) on the other hand, is an apparent-time corpus of Raleigh (North Carolina) English comprising only sociolinguistic interviews.

All SPADE datasets, public or private, are taken to have a *Data Guardian* (DG), i.e. an entity, or often, an individual, with particular responsibility for one or more speech dataset, which they have either collected personally for a specific purpose, overseen the collection of, or now curate (e.g. student projects, inherited collections of speech recordings).

A crucial aspect of the SPADE project has been working closely with our Data Guardians to ensure *best ethical practice* for data sharing (see Chapter 4, this volume). Here, data sharing refers to two kinds of data: (1) primary data, the raw audio + text files, given by the DGs to SPADE for processing (mainly using ISCAN, but sometimes also for forced alignment, to extend dialect coverage); (2) secondary data, anonymized derived datasets of acoustic measures and linguistic data produced by ISCAN for each speech dataset, which SPADE then shares back with each Data Guardian, and – depending on DG permissions – SPADE uses for project research and publications, and/or for deposit for future speech and language research. The secondary datasets are not subject to the same level of scrutiny, since processing with ISCAN renders the primary data anonymous, but ethical practice but must still be considered.

Key *ethical* issues for SPADE relate to (a) the sharing of the primary datasets, since these were usually non-anonymous speech datasets, and hence included personal data, and sometimes also sensitive personal data (i.e. in content, and/or specific metadata for e.g. ethnicity, social practices), and (b) the purpose of the data sharing, specified by each DG. Other important issues arise from SPADE working in compliance with the General Data Protection Regulation (GDPR; legally applicable in the UK/EU from May 2018): (c) SPADE must work with shared primary data according to the specific requirements of each DG (and not with a blanket set of requirements proposed by SPADE). In addition, (d) both 'data holders' (in this case the SPADE team members) and 'data collectors' (the DGs) have to accept responsibility for the received (shared) data, and the data which they share with the project respectively. This means that SPADE has to ascertain, as far as is possible, that the DG is sharing their dataset according to the wishes of the original participants, and that they pass on any specific participants' requirements. So, for example, one speech dataset can be acoustically analysed, but the sound files may only be listened to by SPADE team members who were also members of the original research project team for that dataset. In return, DGs have to respect best ethical practice in their sharing of the secondary derived dataset provided to them by SPADE (see Section 2.3 below).

We managed this aspect of data sharing through a primary/secondary data sharing agreement, which we call the *Data Transfer Agreement* (DTA; cf. Chapter 9, this volume). The DTA was drawn up in conjunction with the Contracts team at the university of the lead PI (Glasgow). The DTA not only confirms the responsibilities of data holders/data collectors, but it also allows DGs to specify what SPADE may (not) do with their data. Data processing for developing the ISCAN software was the minimum, but we also sought additional permission to use acoustic measures in SPADE research and outreach in different ways (e.g. publication of results, using anonymized sound extracts in presentations), and to deposit the derived datasets in a public repository. Through the DTA, DGs have agreed to all, some, or none of the additional data sharing purposes, for all or some of their datasets. For example, one DG allowed the use of anonymized sound extracts from all but nine speakers, as these individuals had not given permission for future use of their audio in this way. At the same time, the same dataset allowed the use of acoustic measures from all speakers for ISCAN development.

Using the DTA, DGs have also indicated specific requirements, such as the exclusion of person and place names from analysis. For example, including the name of a tiny village in Scotland in a derived dataset might lead to identification. We meet this requirement by minimally 'whitelisting' derived datasets: anonymizing all words which are (1) <u>not</u> listed in large electronic English lexicons (Subtlex-US, UK: Brysbaert and New 2009; Van Heuven et al. 2014); and (2) not marked as possible person/place names in the lexicons. Other DG specifications include: citation of a representative publication for their dataset, quotation of grant

numbers, right of veto by the DG for the use of their dataset for particular features, and so on. The main point here is that the DTA gives our DGs free rein to express their wishes, and then our procedures allow for these to be legally checked. All DGs in the British Isles and Canada were offered the opportunity to complete and sign the DTA, with the recommendation that DGs have the agreement signed by their institution on their behalf.

That the DTA could be drawn up for an international project like SPADE, consisting of institutions based in the UK and Canada which share EU data protection laws, and the US, which does not, rested on two underlying agreements. The first was a Research Collaboration Agreement between all participating institutions in all three countries, providing the basis for data sharing within the SPADE project team, especially for the secondary datasets. The second was a Data Sharing Agreement drawn up between the UK and Canadian institutions to specify the basis for primary dataset sharing; the US institutions were not able to participate in this second agreement. This means that the UK and Canadian teams cannot share primary data with the US teams, even that collected from the US itself, since by law, primary data which enters the UK/EU, becomes subject to UK/EU GDPR (so the UK cannot return US primary data to the US). As a result, SPADE observes the following workflow for collecting and sharing primary data:

- The main project site for software development is in Canada. This is also where the master dataset repository is based.
- The UK (Glasgow) team collect primary data from British and Canadian DGs, and then share these with the Canadian team for data analysis using ISCAN. The UK team may also store and carry out data analysis of British and Canadian primary datasets.
- Only secondary, anonymized datasets, from British and Canadian DGs, can be shared with US teams.
- US teams collect primary data from US DGs, and pass them directly to Canada for data processing (and/or process them themselves).

To ensure software development could begin, data collection took place in two phases. Most investigators for SPADE are also DGs, so Phase 1 involved the collection of private datasets held by team members, specifically the Raleigh and SoTC corpora, and four key public corpora (Buckeye, Santa Barbara, ICE-Canada, and SCOTS). The sibilants study (Stuart-Smith et al. 2019a) is based on the dialects from these six corpora.[1] Phase 2 of data collection could only begin after the key data sharing agreements had been drawn up; no GDPR-compliant agreements already existed for us to adapt. The process was fairly lengthy given that the agreements were written while GDPR was coming into effect in the UK. However, our experience with the DTA has been positive, and we are happy to share the documentation and experience with others embarking on large-scale speech data sharing projects such as SPADE. Whilst the details may differ for different jurisdictions, many similar issues may apply. For example, our DTA formed the basis for the agreement used for the US data collection. In terms of procedure, no primary dataset transfer took place before the DTA was agreed, assuming the DG wanted to take up the DTA. Not all DGs wanted to, nor were obliged to, particularly for public datasets or those consisting only of read passages. Once datasets were received, they were checked and cleaned, before passing to the Canadian master repository.

[1] Materials for this study are archived in an OSF project (Stuart-Smith et al. 2019b).

## *2.3. Data citation and acknowledgement*

Collecting a sociolinguistic corpus can be a substantial process which involves designing the sample, recruiting participants, interviewing/recording participants, and collating and transcribing the audio files. Not only do the wishes and rights of the original participants need to be respected by any future user (see Section 2.2 above), we felt that it was essential that the researcher(s) who collected the corpora be given appropriate credit for their hidden labour, which can easily be overlooked. We also wanted to help set a precedent for future data sharing projects of this kind.

Our solution has been to adapt the following convention. The 'SPADE Consortium' appears as the last author for all outputs which use private project-external datasets collected beyond Phase 1. This co-authorship recognizes that the SPADE project, and especially the development of the ISCAN software, would not have been possible without the DGs who generously agreed to share their corpora with us. Their input has been so crucial that we consider the DGs collectively as co-authors of all SPADE-related outputs which make use of private, project-external, primary datasets. Listing all DGs as authors is impractical, and we therefore group them into the 'SPADE Consortium'. In so doing we have followed many of the conventions adopted by the *Atlas of Pidgin and Creole Language Structures* (Michaelis et al. 2013). This convention is adopted for all outputs, irrespective of which corpora are used as the basis of a particular analysis for presentation/publication. The detailed list of the members constituting the SPADE Consortium is given on the project website.

This means that citation of SPADE primary datasets is as follows. Preliminary outputs based on Phase 1 acknowledged the private DGs (SoTC/Raleigh corpora) by co-authorship. All subsequent outputs which use primary or secondary data from SPADE must include the 'SPADE Consortium' as last co-author, and give formal references for the specific corpora used for that output within the text/bibliography of that output. This citation requirement applies to project members, and to all who use the subsequently deposited secondary derived datasets. In this way SPADE DGs can themselves track the future use of secondary data which their data sharing made possible. It especially enables reporting to funders for impact of their research, which is increasingly required in the UK.

## *2.4. Data archiving*

The final core aspect of data sharing for SPADE is the responsible archiving of the primary and secondary datasets, and the databases (in 'Polyglot' format; Section 3). This entails both adhering to the use of standard file formats, and actual primary and secondary dataset storage (Chapters 5 and 7, this volume).

The SPADE primary audio data are in standard file formats (e.g. WAV), usable on any computer. The speech dataset text files are in various human-readable formats as per their deposit. The databases are in a hybrid database format, Polyglot, used by ISCAN (Section 3.2.1). The secondary derived datasets of linguistic data and acoustic phonetic measures are comma-delimited files (.csv), which can be opened in R, Excel, etc.

Project data storage is currently on servers at the Canadian and UK institutions (McGill, Glasgow University). Some US data is also stored at the US sites (North Carolina State University, University of Oregon), respecting ethical issues discussed in Section 2.2. Storage is ensured for at least 10 years and likely for many years afterwards.

Secondary datasets containing the derived linguistic and acoustic measures from all public datasets, and all primary private datasets for which permission has been given, will be

deposited in one or more public data repositories. Example repositories are the Tromso Repository of Language and Linguistics, the Open Science Foundation, or the Linguistic Data Consortium. Secondary datasets (CSVs) will be deposited along with documentation describing their contents and how they were generated using ISCAN (e.g. algorithm parameter values). Such repositories provide sustainable storage of the secondary data in perpetuity, at no or 'incremental cost' to users, and will ensure backup and migration to new formats over time.

# 3. Data processing in SPADE: Integrated Speech Corpus ANalysis ISCAN software

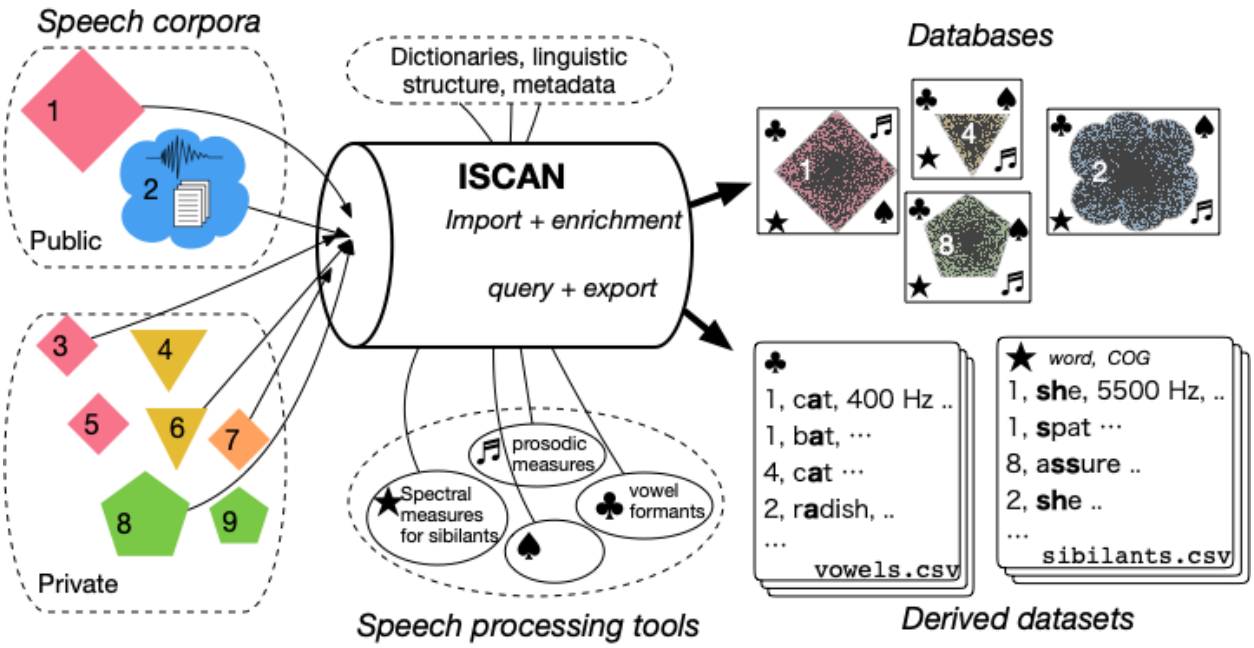


**Figure 3. Schematic of the Integrated Speech ANalysis (ISCAN) system used for data management and processing in SPADE.**

Figure 3 shows the overall data processing pipeline in SPADE. In Section 2 we discussed the initial and final stages of data management for the project: collection of the primary raw speech datasets and sharing of the secondary derived datasets. The steps in between are carried out using software developed for the core purpose of SPADE, scaling up phonetic investigations by carrying out the 'same study' across many speech corpora. The Integrated Speech Corpus ANalysis (ISCAN) software can be configured for different types of cross-corpus analysis. This section describes the goals and design of ISCAN for this specific use case, to show how data processing, storage, access, and search are intended to work for cross-corpus phonetic analyses. Major development of ISCAN is nearly complete. Its implementation is described in detail at iscan.readthedocs.io and polyglotdb.readthedocs.io (see also McAuliffe et al. 2019).

## *3.1. Design goals*

The SPADE use case motivates a number of design goals which hold for any 'big data' project using multiple speech corpora to study linguistic structure:

1. *Scalability*: Speech corpora can be large (1-50 GB each); even basic speech processing algorithms (e.g. pitch extraction) can be slow when run on hours of audio. The system must run in reasonable time as the amount of data grows (the SPADE target corpora contain over 2000 hours of speech).
2. *Abstraction away from corpus format*: Speech corpora are heterogeneous, with numerous formats used over the past 25 years to store annotations and metadata. Extensive scripting is required to perform similar operations on different corpora, despite substantial structural similarities across speech corpora. Users should be able to interact with corpora without understanding particularities of format.
3. *Minimization of technical skill and effort*: Ideally users should need minimal technical skill in order to manipulate and measure acoustic data, and even technical users should be able to minimize scripting by using a standard toolkit.

Each goal also applies for use cases addressed by existing speech data management systems – especially EMU-SDMS, SLAAP, LaBB-CAT, and Phon, whose experiences have helped inform ISCAN development. Further design goals are motivated by SPADE itself, which hold for some but not all cross-corpus phonetic studies:

4. *Enabling multiple users, both local and remote*, to interact with the same dataset - to account for use cases of a single user analyzing their own data (e.g. a DG), and of multiple users at remote locations analyzing the same dataset. Both are needed for SPADE, where team members are in three countries.
5. *Working with restricted datasets*: Many speech datasets cannot be shared or even listened to by groups beyond the original research team (because speech inherently identifies speaker identity). However, neither is in principle necessary for many common phonetic analyses. It should be possible to carry out phonetic analysis on a dataset without access to the raw data, and also for different users to have different levels of access.
6. *Limited functionality to examine and modify individual tokens:* The original vision for SPADE was fully automated analysis without user inspection (Section 1). However, during development we decided to include a limited post-analysis 'inspection' capability (Section 3.3.3), crucially restricted by user permissions.

A schematic of the system is shown in Figure 3. A user goes from raw primary data to derived secondary dataset by:

- importing raw data into a common database format ('import')
- adding linguistic structures and standardized measures to the database, using external speech processing tools, resources such as pronunciation lexicons, and internal algorithms ('enrich')
- then finding relevant tokens ('query') and writing information about them to a CSV file ('export').

(The optional step of 'inspection' is not shown.) These steps can be carried out using either a Python API or a Web GUI (written in Django/AngularJS). Here we mostly abstract away from which interface is used, but assume that the reader would use the GUI.

## *3.2. Data processing and storage*

The first step in processing a raw speech corpus is to import into a standardized database format, meeting the goal of abstracting away from corpus format. ISCAN assumes that minimally phone- and word-level time alignments exist (e.g. Praat 'word', 'phone' tiers), such as the output of a forced aligner. ISCAN can currently import from various TextGrid-based forced aligners (MFA, FAVE, LaBB-CAT) as well as BAS Partitur (used for MAUS: Schiel et al. 1998) and various idiosyncratic corpus-specific formats (TIMIT, Buckeye: Garofolo et al. 1993).

For example, for the /s/-retraction study, six heterogeneous speech corpora were imported: one with an idiosyncratic format (Buckeye) in annotation text files, one in LaBB-CAT format, and several in TextGrids from different forced aligners.

### 3.2.1. Database structure and 'Import'

Use of a database presupposes a data model. There are two parts to ISCAN's data model, corresponding to the structure of transcribed speech and common measures:

- *Annotation graphs* (Bird and Liberman 2001; also used in the EMU-SDMS, LaBB-CAT systems): a formalism based on graphs (in the sense of nodes and edges) which captures the logical structure underlying transcribed speech. Nodes and edges in the graph represent points in time, and intervals of time over which an annotation occurs (e.g. /k/ in the word *cat*).
- *Time series*: acoustic measures defined at fixed intervals over time, such as an f0 track.

Our corresponding custom database format, called *Polyglot*, uses two sub-databases, each matched to the structure of one aspect of the data:

- Neo4j, a NoSQL *graph database*, is used to represent transcribed speech via annotation graphs, which capture linguistic objects and their temporal relationships. For example, in the initial import, word and phone information are parsed into a meaningful structure, reflecting both hierarchical information (e.g. which phones belong to which word; which phone follows which) and type-token relationships (such as what properties are shared across all productions and which words/phones are spoken by a particular speaker).
- InfluxDB, a NoSQL *time-series database*, stores time series associated with a particular token of a linguistic object (e.g. the f0 track across a word).

The 'polyglot persistence' design of our system, where different sub-databases are used for different data types, should maximize scalability, one of our key goals, because each sub-database is already optimized for the structure of a particular data type. Our choice of a graph database in particular over a relational database (used in e.g. EMU, LaBB-CAT, and SLAAP) was motivated by scalability. A disadvantage of this choice is the high storage footprint associated with graph databases. Empirically, for SPADE we have found that after enough enrichment is performed (see below) to do a typical phonetic study, the resulting Polyglot database is about as large as the original speech corpus.

### 3.2.2. Processing and storage loop: 'enrichment'

The database resulting from importing a corpus is of limited use for phonetic studies – only word or phone durations can be examined. For a typical use case, the database is first built up through

'enrichment': a loop of data processing and storage, to add different linguistic objects and phonetic measures of several types:

- *New linguistic units* can be created, to enhance the structured hierarchy representing the corpus in the database. For example, words can be grouped together into larger chunks ('utterances'), or phones into syllables.
- *Non-acoustic properties* can be added to linguistic objects, including properties of words/phones/speakers, from external resources such as pronunciation lexicons (e.g. syllable stress, word frequency) or corpus metadata files (e.g. speaker gender, age), and measures based on hierarchical relationships – such as speech rate (e.g. 'syllables per second') or number of phones in a word.
- *Acoustic measures* can be stored, by processing the raw sound files using internal algorithms or integration with external tools (such as AutoVOT or Reaper: Talkin 2015). Currently available measures include f0, vowel formants (algorithm described in Mielke et al. 2019), and voice onset time, as well as anything computable by a user-specified Praat script in a certain format. Continuous-time measurements (such as f0) can be stored as single points (e.g. one f0 per vowel) or tracks (e.g. one f0 track per vowel).

For example, for the /s/-retraction study, enrichment included:

- Phone position in word, syllables, stress (from an external lexicon)
- Speaker dialect and gender (from corpus metadata)
- Acoustic measures for each sibilant token, such as Center of Gravity and spectral slope, calculated by a user-specified Praat script

Enrichment is a loop because all new information computed is stored, and subsequent enrichment steps usually depend on previously stored information. Anything encoded in enrichment is stored in the database, and can be used again in the future. This design choice follows from the intended workflow of ISCAN: data processing and storage are only done once, can be slow, and require access to the original speech data. However, once the database is created, it exists independently of the original speech data, and can be used efficiently in different studies: querying the database and writing the results to a data file are designed to be fast.

## *3.3. Interacting with a Polyglot database: access, search, inspection*

Importing a speech corpus and carrying out enrichment results in one *Polyglot database* per corpus (top right of Figure 3). Each database contains all information needed to carry out common phonetic analyses without access to the raw corpus itself. We now describe how users interact with these databases, in terms of *access*, *searching*, and *inspection*.

### 3.3.1. Access: system configuration and user access

For our use case of cross-corpus analysis, we presume many existing Polyglot databases corresponding to different datasets; these need to be accessed by multiple users, possibly at different sites (goal #4), with access restrictions that respect ethical restrictions on some datasets (goal #5). Access to and interaction with the databases is managed via a Django web framework. The overall ISCAN system, which consists of several pieces managed by Docker, is called the *ISCAN server*.

An ISCAN server is installed on a static machine (e.g. a desktop), from where it can be used to interact with the databases locally by its installer. If the machine is a web server, the databases can also be accessed by remote users – which is the case we assume for the rest of this section, and the typical use case for SPADE. A fully-fledged permissions system allows a user's access and functionality to be restricted for particular databases.[2] This is most relevant for 'inspection' (Section 3.3.3) to disable users' ability to listen to audio or see identifying information (e.g. parts of the transcript) for a dataset. However, the permissions system could in principle be used more generally – such as to enable only searching for tokens but not exporting a data file, if a researcher wanted to let others explore the data they used for a study, but is not able to provide a readable CSV, as for example, during a class.

To give a concrete example: an ISCAN server for the SPADE project has been set up on a web server at McGill, where users can log in by going to a web address. A user who has been given a tutorial account would only see the 'iscan-tutorial' database (a subset of the ICE-Canada corpus), and would not see some functionality which has been disabled for purposes of the tutorial. A SPADE team member who logs in would see all available databases, and all functionality.

3.3.2. Searching: 'query' and 'export'

However the ISCAN server is accessed (either locally or on a remote web server), a user primarily interacts with a Polyglot database by executing *queries*, to find a subset of linguistic objects (e.g. phones, words) of interest for a phonetic study. Queries are constructed either in a graphical interface (in the GUI: Figure 4) or in a custom Python query language (in the Python API) – no knowledge of underlying query languages of the sub-databases is assumed. Using a custom query language (as also in EMU-SDMS) minimizes the technical knowledge required from the user. Queries can reference properties of an annotation (e.g. 'all phones which are labeled /s/ or /ʃ/'), user-defined subsets (e.g. 'sibilants'), information from associated linguistic objects (e.g. following phone, syllable stress), and so on. For example, for the /s/-retraction study, the query found all word-initial sibilants in stressed syllables. Once the subset of objects has been found, information about them can be exported to a data file (a CSV), with one row per token and one column per variable. Any information referenced above (in query, enrichment, import) can be exported as a column, including acoustic measures. For example, in the /s/-retraction study, the exported data file included information about any following phones in the syllable onset, speaker demographics (dialect, gender), and acoustic measures associated with each sibilant (e.g. spectral Center of Gravity, peak, etc.).

[2] Such permissions systems are important for some other speech database management systems, such as SLAAP, which hosts mostly restricted datasets.

**Figure 4. Screenshot of 'query interface' for ISCAN.**

### 3.3.3. Inspection

The original vision of SPADE envisaged fully automatic analysis of speech, without user access to raw corpora. As ISCAN was developed and used team-internally, it became clear that *fully*-automatic analysis is not yet realistic in many practical settings, such as when developing a new analysis pipeline to be applied across many corpora. For any phonetic study, even large-scale automated ones, some inspection of individual tokens may be important to get a handle on the data, and to examine unusual cases. The ISCAN GUI thus contains functionality for 'inspection'. Subject to user permissions, tokens returned by a query can be visually and auditorily inspected using a waveform, spectrogram, and a TextGrid-like display showing the linguistic context (phones, words) (Figure 5).

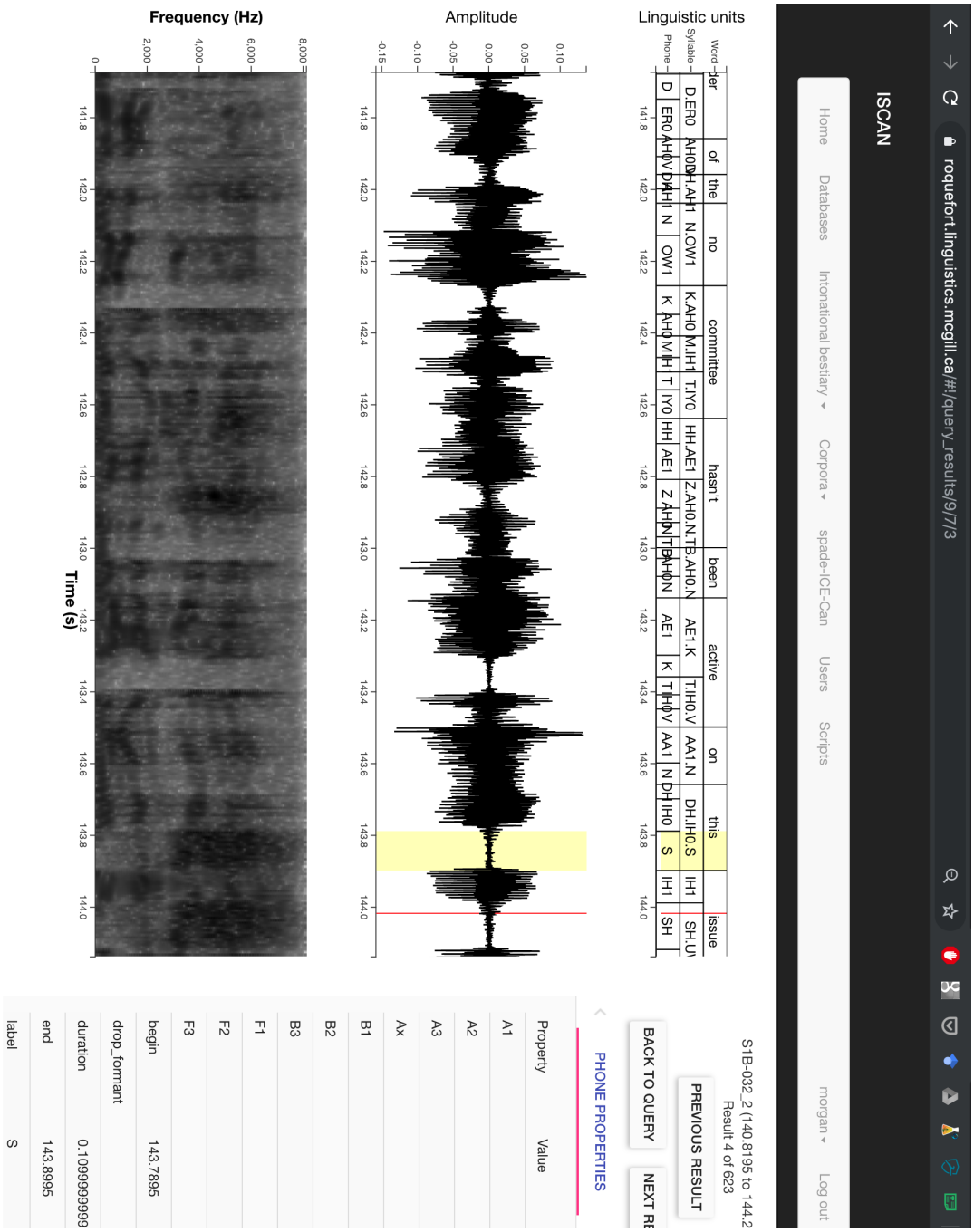


**Figure 5. Screenshot of 'inspection interface' for ISCAN.**

Importantly, this functionality is controlled by the permissions system: access to each aspect is defined on a per-corpus and per-user basis. For example, users' ability to actually play audio for individual tokens can be disabled – which addresses the primary privacy concern associated with sharing speech data. In this way, examination of individual tokens should be possible even for (some) private datasets.

ISCAN contains limited functionality to allow the user to change the database through Inspection, in two ways which are important for augmenting 'automatic' analysis. Tokens can be *excluded* from further analysis (e.g. due to a bad f0 track or incorrect forced alignment), and annotations can be *corrected*, though currently the latter is restricted to fixing f0 tracks. While more extensive Inspection capability is important for enabling large-scale studies, this direction is beyond the remit of the SPADE project. Inspection functionality is better developed in other speech database management systems (e.g. LaBB-CAT, Phon, EMU-SDMS) whose intended use cases centrally involve corpus annotation.

# 4. Future directions for ISCAN and SPADE

The SPADE project is a concrete instantiation, and first step, of developing our philosophy of speech data management for analyses of multiple speech corpora, where corpora are translated into standardized databases from which consistent high-quality acoustic measures can easily be extracted for analysis. The SPADE use case necessarily focuses on subsets of speech segments in a subset of English dialects. Obvious extensions are to extend our sample to include overseas, native and non-native Englishes, and to extend to other languages for cross-linguistic research (Sonderegger et al. 2017). We also need to adapt ISCAN for large-scale study of suprasegmental speech phenomena, building on the parallel development of the system for smaller-scale cross-corpus analyses in the Intonational Bestiary project (Goodhue et al. 2016), as described in McAuliffe et al. (2019).

To this end, we also regard ISCAN as a step in a continually developing software system for integrated speech corpus analysis (ISCAN after all, continues Speech Corpus Tools: McAuliffe et al. 2017b). In order to address longevity of the system, all code is freely available and open source (on Github repositories: MontrealCorpusTools/ISCAN, MontrealCorpusTools/iscan-spade-server, MontrealCorpusTools/PolyglotDB), including stable releases. We hope and expect that the code will be taken up, extended, and cannibalized in the future, if not by ourselves, then by others who are also working towards the general scientific direction of scaling up phonetic data analysis.

Acknowledgements

The research reported here was funded through the 4th Digging into Data Challenge issued by the Trans-Atlantic Platform, through grants ESRC ES/R003963/1, NSERC/CRSNG RGPDD 501771-16, SSHRC/CRSH 869-2016-0006, and NSF 1730479. We thank all non-author SPADE team members for their contributions, especially Jeff Mielke, Robin Dodsworth, Erik Thomas, Vanna Willerton, James Tanner, Michael Goodale, and Arlie Coles.

# References


Anderson, Jean, David Beavan, and Christian Kay. 2007. SCOTS: Scottish Corpus of Texts and Speech. In *Creating and Digitizing Language Corpora*, edited by Joan Beal, Karen Corrigan, and Hermann Moisl, 17–34. London: Palgrave Macmillan.

Bird, Steven, and Mark Liberman. 2001. A Formal Framework for Linguistic Annotation. *Speech Communication* 33 (1): 23–60.

Boersma, Paul, and David Weenink. 2016. Praat: Doing Phonetics by Computer (Version 6.0.19). http://www.praat.org/.

Brysbaert, Marc, and Boris New. 2009. Moving Beyond Kučera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English. *Behavior Research Methods* 41 (4): 977-990.

Coleman, John, Ladan Baghai-Ravary, John Pybus, and Sergio Grau. 2012. Audio BNC: The Audio Edition of the Spoken British National Corpus. Phonetics Laboratory, University of Oxford. http://www.phon.ox.ac.uk/AudioBNC.

Dodsworth, Robin, and Mary Kohn. 2012. Urban Rejection of the Vernacular: The SVS Undone. *Language Variation and Change* 24 (2): 221–245.

Du Bois, John, Wallace Chafe, Charles Meyer, Sandra Thompson, Robert Englebretson, and Nii Martey. 2000-2005. Santa Barbara Corpus of Spoken American English, Parts 1-4. Philadelphia: Linguistic Data Consortium.

Fromont, Robert, and Jennifer Hay. 2012. LaBB-CAT: An Annotation Store. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, edited by Paul Cook and Scott Nowson, 113-117.

Garofolo, John, Lori Lamel, William Fisher, Jonathan Fiscus, David Pallett, Nancy Dahlgren, and Victor Zue. 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Philadelphia: Linguistic Data Consortium.

Goodhue, Daniel, Lyana Harrison, Y. T. Clementine Su, and Michael Wagner. 2016. Toward a Bestiary of English Intonational Tunes. In *Proceedings of the 46th Conference of the North Eastern Linguistic Society (NELS)*, edited by Christopher Hammerly and Brandon Prickett, 311-320.

Grabe, Esther, Brechtje Post, and Francis Nolan. 2001. The IViE Corpus. Department of Linguistics, University of Cambridge. http://www.phon.ox.ac.uk/IViE.

Kendall, Tyler. 2007. Enhancing Sociolinguistic Data Collections: The North Carolina Sociolinguistic Archive and Analysis Project. *Penn Working Papers in Linguistics* 13 (2), 15-26. Philadelphia: University of Pennsylvania.

Keshet, Joseph, Morgan Sonderegger, and Thea Knowles. 2014. AutoVOT: A Tool for Automatic Measurement of Voice Onset Time Using Discriminative Structured Prediction (Version 0.91). https://github.com/mlml/autovot/.

Kisler, Thomas, Florian Schiel, and Han Sloetjes. 2012. Signal Processing via Web Services: The Use Case WebMAUS. In *Proceedings of the Digital Humanities Conference 2012*, edited by Clare Mills, Michael Pidd, and Esther Ward, 30-38. Sheffield, UK: HRI Online Publications.

Liberman, Mark. 2019. Corpus Phonetics. *Annual Review of Linguistics* 5: 91-107.

McAuliffe, Michael, Arlie Coles, Michael Goodale, Sarah Mihuc, Michael Wagner, Jane Stuart-Smith, and Morgan Sonderegger. 2019. ISCAN: A System for Integrated Phonetic Analyses Across Speech Corpora. In *Proceedings of the 19th International Congress of Phonetic Sciences*, 1322-1326.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017a. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proceedings of Interspeech 2017,* 498–502.

McAuliffe, Michael, Elias Stengel-Eskin, Michaela Socolof, and Morgan Sonderegger. 2017b. Polyglot and Speech Corpus Tools: A System for Representing, Integrating, and Querying Speech Corpora. In *Proceedings of Interspeech 2017*, 3887–3891.

Michaelis, Susanne Maria, Philippe Maurer, Martin Haspelmath, and Magnus Huber, eds. 2013. *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://apics-online.info.

Mielke, Jeff, Erik Thomas, Josef Fruehwald, Michael McAuliffe, Morgan Sonderegger, Jane Stuart-Smith, and Robin Dodsworth. 2019. Age Vectors vs. Axes of Intraspeaker Variation in Vowel Formants Measured Automatically from Several English Speech Corpora. In *Proceedings of the 19th International Congress of Phonetic Sciences*, 1258-1262.

Newman, John, and Columbus, Georgia. 2010. The ICE-Canada Corpus (Version 1). University of Alberta. https://dataverse.library.ualberta.ca/dataverse/VOICE.

Nolan, Francis, Kirsty McDougall, Gea de Jong, and Toby Hudson. 2009. The DyViS database: Style-Controlled Recordings of 100 Homogeneous Speakers for Forensic Phonetic Research. *International Journal of Speech, Language and the Law* 16 (1): 31-57.

Pitt, Mark, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. 2007. Buckeye Corpus of Conversational Speech (2nd release). Columbus: Department of Psychology, Ohio State University.

Rose, Yvan, Brian MacWhinney, Rodrigue Byrne, Gregory Hedlund, Keith Maddocks, Philip O'Brien, and Todd Wareham. 2006. Introducing Phon: A Software Solution for the Study of Phonological Acquisition. In *Proceedings of the 30th Annual Boston University Conference on Language Development*, 489-500.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard, and Jiahong Yuan. 2015. FAVE (Forced Alignment and Vowel Extraction) Program Suite (version 1.2.2).

Schiel, Florian, Susanne Burger, Anja Geumann, and Karl Weilhammer. 1998. The Partitur Format at BAS. In *Proceedings of the First International Conference on Language Resources and Evaluation*, 1295–1301.

Sonderegger, Morgan, Michael McAuliffe, and Hye-Young Bang. 2017. Segmental Influences on F0: Cross-Linguistic and Interspeaker Variability of Phonetic Precursors. Paper presented at the 4th Workshop on Sound Change, Edinburgh, April 20-22.

Stuart-Smith, Jane, Morgan Sonderegger, Rachel Macdonald, Jeff Mielke, Michael McAuliffe, and Erik Thomas. 2019a. Large-Scale Acoustic Analysis of Dialectal and Social Factors in English /s/-Retraction. In *Proceedings of the 19th International Congress of Phonetic Sciences*, 1273-1277.

Stuart-Smith, Jane, Morgan Sonderegger, Rachel MacDonald, Jeff Mielke, Michael McAuliffe, and Erik Thomas. 2019b. Large-Scale Analyses of English /s/-Retraction Across Dialects. OSF. osf.io/bknrg.

Stuart-Smith, Jane, Brian José, Tamara Rathcke, Rachel Macdonald, and Erik Lawson. 2017. Changing Sounds in a Changing City: An Acoustic Phonetic Investigation of Real-Time Change over a Century of Glaswegian. In *Language and a Sense of Place: Studies in Language and Region*, edited by Chris Montgomery and Emma Moore, 38-65. Cambridge: Cambridge University Press.

Talkin, David. 2015. REAPER: Robust Epoch And Pitch EstimatoR. https://github.com/google/REAPER.

Thomas, Erik. 2001. *An Acoustic Analysis of Vowel Variation in New World English*. Publication of the American Dialect Society 85. Duke University Press.

Van Heuven, Walter J. B., Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology* 67 (6): 1176-1190.

Winkelmann, Raphael, Jonathan Harrington, and Klaus Jänsch. 2017. EMU-SDMS: Advanced Speech Database Management and Analysis in R. *Computer Speech & Language* 45: 392–410.