ISCAN: A SYSTEM FOR INTEGRATED PHONETIC ANALYSES ACROSS SPEECH CORPORA

Michael McAuliffe[a], Arlie Coles[a], Michael Goodale[a], Sarah Mihuc[a], Michael Wagner[a], Jane Stuart-Smith[b], Morgan Sonderegger[a]

[a]McGill University, Canada; [b]University of Glasgow, United Kingdom
{michael.mcauliffe, arlie.coles, michael.goodale, sarah.mihuc}@mail.mcgill.ca, chael@mcgill.ca, jane.stuart-smith@glasgow.ac.uk, morgan.sonderegger@mcgill.ca

ABSTRACT

Speech corpora of many languages, styles, and formats exist in the world, representing significant potential for the phonetic sciences. However in practice there are significant practical and methodological barriers to conducting the "same study" across corpora, including necessary technical skills and non-comparability of results using non-standardized measures. We introduce an open-source software system for Integrated Speech Corpus ANalysis (ISCAN), which enables automated acoustic phonetic analysis across spoken corpora of diverse formats and sizes. A web-browser-based GUI and Python package allow for different user backgrounds. The system is a major update of core functionality for fully- automated speech corpus analysis (importing, enriching, querying) from a previous version, to meet new goals: different user configurations, working with restricted datasets, and interacting with data (visualization and correction). The system's flexibility for different projects is shown in two use cases: large-scale automatic segmental analysis of spontaneous speech across English dialects, and smaller-scale semi-automatic prosodic analysis.

Keywords: speech corpora, speech analysis software, large-scale studies, speech technology

1. INTRODUCTION

A huge number of speech corpora annotated at least with an orthographic transcription exist, from small-scale collections of controlled laboratory speech to large-scale spontaneous speech datasets At the same time, increasingly accurate tools exist for aligning speech [7, 9, 11, 17] and automatically measuring variables used in phonetic research (e.g. formants, f0, VOT: [8, 17, 21]). This confluence of data and methods means that it is now possible to scale up phonetic research, with obvious scientific potential [10, 24], via *integrated corpus analysis*: carrying out similar studies across many different corpora.

In practice, there are serious practical barriers to corpus phonetics, which motivate the development of systems to facilitate phonetic analyses across corpora. Such a system should meet several goals:

- *Scalability*: Speech corpora can be large, so cross-corpus analyses require substantial storage and computation resources. The system must perform in reasonable time as the amount of data grows.
- *Abstraction away from corpus format*: Speech corpora are complex and heterogeneous, with metadata and annotation files in many different formats. Users should be able to interact with corpora without understanding their format.
- *Minimize necessary technical skill and effort*: Technical skill is currently required for corpus studies, with researchers writing extensive scripts to perform similar operations on different corpora. Minimal technical skill should be needed from users and technically-skilled users should be able to minimize scripting, by leveraging standardized measures which make analyses across different corpora comparable.

Among several systems developed for management and analysis of speech corpora in recent years [2, 4, 16, 19, 23], the Polyglot-Speech Corpus Tools (SCT) system [12] was optimized for large-scale studies across many corpora, prioritizing the goals above. This system enabled fully-automated corpus studies by importing speech datasets into a common database format, enriching each database with standardized measures, finding relevant tokens, and exporting a data file. A Python API and desktop GUI allowed for different user skill levels, and the intended use case was a single research group, analyzing corpora to which users had unrestricted access.

This paper describes a major update to this system, ISCAN (**I**ntegrated **S**peech **C**orpus **AN**alysis), motivated by additional goals which aim to broaden the types of corpus study the system can be used for:

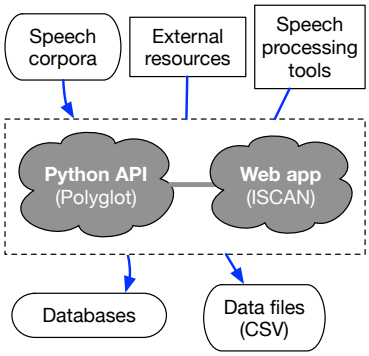

Figure 1: System architecture: software (center) and elements used (top) or created (bottom) by the software.

- *Different user configurations* should be possible, from a single user/group analyzing their own data to multiple users at different locations analyzing the same data. More generally, the system should be configurable for different projects.
- The system should be *usable with restricted datasets*: many speech datasets cannot be listened to (by other groups) or transfered due to ethics concerns—but in principle neither is required for common phonetic analyses.
- *Enable semi-automated analysis*: visualizing, checking, and correcting measures for individual tokens is crucial for working with smaller datasets, and in practice necessary even for large-scale 'automated' studies.

These goals stem from two complementary use cases for which the system was developed. (Sec. 4).

2. SYSTEM ARCHITECTURE

As schematized in Figure 1, users create, interact with, and query databases in a custom format (data representation and storage are described in [12]). The system can be used via a Python API or a web browser graphical user interface (GUI), with the exception of Inspection (visualizing/correcting data), which requires the GUI. This architecture shares broad similarities with other speech database management systems, especially EMU, LaBBCAT, and Phon [23, 4, 16]: server-client architecture, database storage, and external tool integration.

The Python API is conceptually similar to Polyglot-SCT; we focus on the web app (ISCAN), which contains the GUI, and which is completely new. ISCAN is developed in Django and AngularJS,[1] flexible and widely-used web frameworks which enable meeting the goals of working with restricted data and flexible user configurations. A fully-fledged permissions system allows user access to any functionality to be disabled for a particular dataset—such as a user being able to query but not inspect, for data which must remain anonymous. A server-client architecture separates users (the client web browser) from raw data/databases (the server), enabling either multiple users interacting with the same database (remotely), or a user(s) working on data on their own laptop (etc.).[2] ISCAN can be configured differently in different installations, to facilitate integrated corpus analysis in the context of different projects. Two configurations have been used, corresponding to the two use cases described below (Sec. 4): `iscan-spade` and `iscan-bestiary`.

3. IMPLEMENTATION AND FEATURES

Using ISCAN, a user goes from raw data to a CSV file via:

1. *Import*: data from speech corpora is stored in a standardized database format.
2. *Enrichment*: various measures (e.g. syllable position, vowel formants) are added to the databases using external speech processing tools and resources and internal algorithms.
3. *Query*: find tokens of interest in the databases, with associated measures.
4. *Inspection* (optional): visualization, editing, and correction of individual tokens.
5. *Export*: save to a data file.

Figure 2 shows a schematic example of these steps in the context of an application using the `iscan-spade` system. Functionality for each step is summarized below, with more details given in ISCAN's documentation and tutorials.[3]

The intended use case for the system is for import and enrichment to be once per corpus, since these steps can be slow, while query/inspection/export are faster, and can be done repeatedly to address different research questions in different studies.

3.1. Import

First, raw data in a speech corpus is imported into a standardized database format, allowing the system to abstract away from different corpus formats (as in Fig. 2: different shape datasets $\rightarrow$ same shape databases). The importer parses raw annotations (e.g. Praat 'phone', 'word' TextGrid tiers) into a meaningful hierarchy (e.g. each 'phone' token belongs to a 'word' token), which greatly facilitates subsequent steps. We assume at least phone and word time-alignments exist, such as the output of various forced aligners [7, 11, 17]. Corpora in LaBB-CAT, BAS Partitur [18], TIMIT, or Buckeye [5, 15] formats are also currently supported.

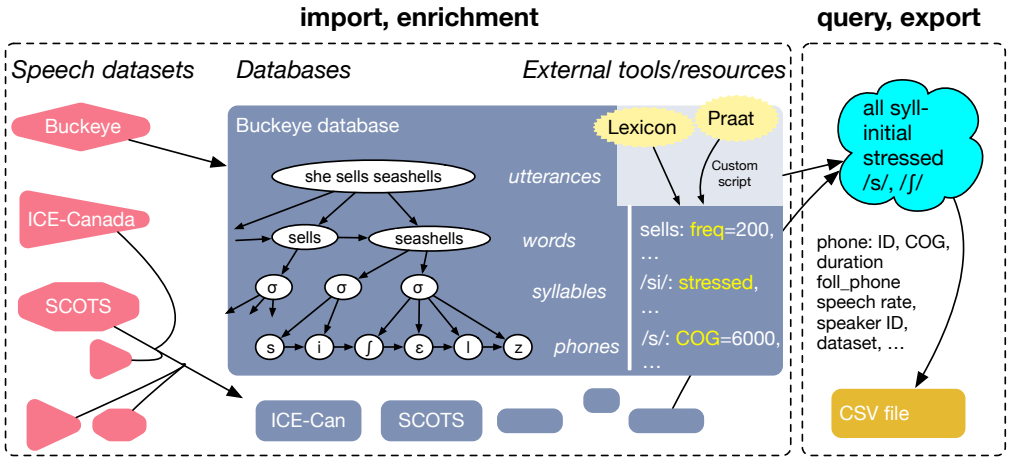

Figure 2: Schematic example of `iscan-spade` use to generate data for a study [20] examining realization of /s/ and /ʃ/ across English dialects.

3.2. Enrichment

Importing a corpus results in a database containing only the word and phone levels (in a hierarchy), plus speaker and sound file IDs. Most of a database's contents result from subsequent *enrichment*, which adds different types of information and measures often used in linguistic studies.

Non-acoustic information: new units can be added to the hierarchy. Words can be grouped into connected-speech chunks ('utterances'), separated by non-speech intervals (e.g. pauses) above a minimum length, and phones into syllables, using the maximum onset algorithm. Measures based on hierarchical relations can be added, such as speech rate (e.g. syllables/second in an utterance) or position (e.g. phone position in a syllable). Properties of words, phones, speakers, or sound files can be added, from external resources such as lexicons (e.g. word frequency, stress pattern), or user-specified speaker metadata (e.g. gender) or annotation files.

Acoustic measures from the raw sound files can be calculated and stored. Currently available measures include f0, intensity, formants, voice onset time (VOT)—and any measure computable by a simple user-uploaded Praat script. The system supports integration with external speech processing tools (currently: Praat, Reaper for f0; AutoVOT for VOT [3, 21, 8]), and contains powerful internal methods, such as the current vowel formant algorithm (described in [13]). Continuous-time acoustic measures (e.g. f0, formants) can be encoded as both points or tracks—for example, one f0 value or f0 track per vowel token—enabling studies using both 'static' and 'dynamic' measures.

Relativized measures: any quantitative measure calculated in enrichment can be normalized in ways useful for phonetic studies. For example, f0 can be scaled relative to a speaker's range, or phone (token) duration computed relative to the phone's average duration in the corpus.

Goals for future development of enrichment functionality include alternative methods for adding non-acoustic information, and standardization of the interface with external tools (to facilitate integration of a broader range of tools).

3.3. Query and Export

Given an enriched database, the user conducts a *query* to find a subset of linguistic objects of interest (e.g. words, phones), and then *exports* information about them. Queries are constructed in either a graphical interface or a custom Python query language—no knowledge of the query languages of the underlying databases (e.g. SQL) is required. Queries can reference aspects of an annotation ("all phones which have label 's"'), user defined subsets (e.g."sibilants"), associated information (e.g. following phones), hierarchical relations, and so on. For each token, this information can be returned as a column of the exported CSV—as well as any acoustic enrichment, such as formant tracks.

3.4. Inspection

For any phonetic study, even large-scale automated studies, getting a handle on the data through inspection of individual tokens is important. ISCAN allows for both visual and auditory inspection. Any token in a query result can be inspected, giving the full context of the utterance (surrounding words/phones,

spectrogram) and audio playback. This allows for problematic tokens (due to alignment errors, etc.) to be caught and excluded before data analysis (from the exported CSV) begins.

Another key feature of the inspection view is the ability to correct the automatically generated data. For instance, pitch algorithms often have octave errors in the pitch track, which can be corrected via a graphical interface. Crucially, access to inspection is defined on a per-corpus, per-user basis, via the permissions system. For example, only certain users can perform corrections or play audio.

4. USE CASES

ISCAN has been developed in the context of two integrated corpus analysis projects. The Speech Across Dialects of English (*SPADE*) project examines larger datasets (5-100+ hours), mostly conversational speech, focusing on segmental realization; the *Intonational Bestiary* project examines smaller datasets from production experiments, focusing on prosody. Polyglot's database structure enables studies of 'corpora' of both sizes: making large-scale studies computationally feasible by efficient organization and storage of rich metadata, while allowing smaller-scale studies to leverage the same organization and metadata. The two projects use different configurations of ISCAN, which essentially means turning off some functionality for a cleaner user experience. For instance, because Intonational Bestiary studies focus on entire intonation contours, most Query and Export functionality at the sub-utterance level is disabled in `iscan-bestiary`.

4.1. ISCAN-SPADE

The `iscan-spade` configuration is used for SPADE, a multi-site project whose remit is large-scale study of spoken English across space (UK, US, Canada) and time.[4] The project is analyzing several dozen datasets; some are publicly available (e.g. Buckeye Corpus), while many are private corpora provided by data guardians, for which ISCAN's permissions system is important. All raw datasets are kept at one project site, where ISCAN is hosted on a web server, providing access to the (anonymous, derived) databases built from these datasets to project team members at other sites via web browser. ISCAN's server-client architecture enables this setup, which respects data protection considerations by separating raw data from users.

`iscan-spade` is being used to carry out large-scale analyses of segmental realization across English dialects—such as the sibilant acoustics study shown in Fig. 2 and [20], where stressed-word-initial sibilants are analyzed to characterize the degree of 'retraction' of /s/, relative to /ʃ/ production, as a function of onset structure. This study includes import from corpora in different formats, from dialects across the US, UK, and Canada (e.g. Buckeye, SCOTS, ICE-Canada: [1, 14]) Enrichment includes acoustic measures characterizing sibilant production, such as center of gravity (COG) and duration, computed via a user-specified Praat script. These measures are included in query and export of a data file, which includes information to address the study's research question (e.g. acoustic measures, following segment info) and controls. Another case study so far analyzes vowel formants [13].

4.2. Bestiary

The `iscan-bestiary` set up has been used to analyze experimental data from production studies focused on prosody, and leverages the linguistic enrichment and structured nature of data in ISCAN. The first project that `iscan-bestiary` has been used for is the Intonational Bestiary [6].[5] ISCAN generates utterance pitch tracks per sound file (enrichment), allows for manual correction of the pitch tracks for octave errors (inspection), and exports the pitch tracks along with other metadata about the experimental conditions and the speakers (export). `iscan-bestiary` also allows for adding sound-file-level annotations, capturing for example which intonational 'tune' was used for an utterance.

Another project using `iscan-bestiary` was an investigation of three functions of prosody (intonational tune, contrastive focus and phrasing) in a production study, similar to [22]. Utterance pitch tracks for each production were automatically generated, then hand-corrected (enrichment, inspection). The words of interest in this study were proper names. From each of these words of interest, max F0 (from the pitch track), mean intensity (from an intensity track), and duration were extracted for each syllable. Thus, the study used dynamic tracks generated and corrected in ISCAN, as well as ISCAN's enriched hierarchical structure of linguistic units.

Acknowledgements

We thank the SPADE team, especially R. Macdonald, J. Tanner, S. Willerton, T. Kendall, and K. Gunter, for testing and feedback. We acknowledge funding from the Trans-Atlantic Platform Digging Into Data program (SSHRC/NSERC/ESRC/NSF); and from SSHRC grant #435-2014-1504 and the SSHRC CRC program to MW.

5. REFERENCES

[1] Anderson, J., Beavan, D., Kay, C. 2007. Scots: Scottish corpus of texts and speech. In: Beal, J., Corrigan, K., Moisl, H., (eds), *Creating and digitizing language corpora*. Springer 17–34.

[2] Bigi, B. 2015. SPPAS - multi-lingual approaches to the automatic annotation of speech. *The Phonetician* 111–112, 54–69.

[3] Boersma, P., Weenink, D. 2017. Praat: doing phonetics by computer [computer program].

[4] Fromont, R., Hay, J. 2012. LaBB-CAT: An annotation store. *Proc. Australasian Language Technology Association Workshop 2012* Dunedin. 113–117.

[5] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N. 1993. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia: Linguistic Data Consortium.

[6] Goodhue, D., Harrison, L., Su, Y. T. C., Wagner, M. 2016. Toward a bestiary of English intonational tunes. Hammerly, C., Prickett, B., (eds), *Proceedings of the 46th Conference of the North Eastern Linguistic Society (NELS)* Montreal. 311–320.

[7] Gorman, K., Howell, J., Wagner, M. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics* 39(3), 192–193.

[8] Keshet, J., Sonderegger, M., Knowles, T. 2014. AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction [Computer program]. Version 0.91. Available at https://github.com/mlml/autovot/.

[9] Kisler, T., Schiel, F., Sloetjes, H. 2012. Signal processing via web services: the use case WebMAUS. *Proc. Digital Humanities Conference 2012* Hamburg.

[10] Liberman, M. 2019. Corpus phonetics. *Annual Review of Linguistics* 5, 15.1–15.17.

[11] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. 2017. Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *Proc. Interspeech 2017* Stockholm. 498–502.

[12] McAuliffe, M., Stengel-Eskin, E., Socolof, M., Sonderegger, M. 2017. Polyglot and Speech Corpus Tools: a system for representing, integrating, and querying speech corpora. *Proceedings of Interspeech 2017* Stockholm. 3887–3891.

[13] Mielke, J., Thomas, E., Fruehwald, J., Stuart-Smith, J., Sonderegger, M., Dodsworth, R., McAuliffe, M. 2019. Age vectors vs. axes of intraspeaker variation in vowel formants measured automatically from several English speech corpora. *Proc. 19th ICPhS* Melbourne.

[14] Newman, J., Columbus, G. 2010. The ICE-Canada Corpus, Version 1. Speech portion available at https://dataverse.library.ualberta.ca/dataverse/VOICE.

[15] Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E. 2007. *Buckeye Corpus of Conversational Speech (2nd release)*. Columbus: Department of Psychology, Ohio State University.

[16] Rose, Y., MacWhinney, B. 2014. The PhonBank Project: Data and software-assisted methods for the study of phonology and phonological development. In: Durand, J., Gut, U., Kristoffersen, G., (eds), *The Oxford Handbook of Corpus Phonology*. Oxford University Press 308–401.

[17] Rosenfelder, I., Fruehwald, J., Evanini, K., Yuan, J. 2011. FAVE (forced alignment and vowel extraction) program suite. http://fave.ling.upenn.edu.

[18] Schiel, F., Burger, S., Geumann, A., Weilhammer, K. 1998. The Partitur format at BAS. *Proc. First International Conference on Language Resources and Evaluation* Grenada. 1295–1301.

[19] Schmidt, T., Wörner, K. 2009. EXMARaLDA–creating, analyzing and sharing spoken language corpora for pragmatics research. *Pragmatics* 19(4), 565–582.

[20] Stuart-Smith, J., Sonderegger, M., Macdonald, R., Mielke, J., McAuliffe, M., Thomas, E. 2019. Large-scale acoustic analysis of dialectal and social factors in English /s/-retraction. *Proc. 19th ICPhS* Melbourne.

[21] Talkin, D. 2015. REAPER: Robust Epoch And Pitch EstimatoR [computer program]. https://github.com/google/REAPER.

[22] Wagner, M., McAuliffe, M. 2017. Three dimensions of sentence prosody and their (non-) interactions. *Proc. Interspeech 2017* Stockholm. 3196–3200.

[23] Winkelmann, R., Harrington, J., Jänsch, K. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language* 45, 392–410.

[24] Yuan, J., Lai, W., Cieri, C., Liberman, M. 2018. Using forced alignment for phonetics research. In: *Chinese Language Resources and Processing: Text, Speech and Language Technology*. Springer.

[1] https://www.djangoproject.com/, https://angularjs.org/

[2] The package is downloaded and installed on a local machine, setting up the ISCAN server. Users can interact with the ISCAN server locally (`localhost`), or remotely if the machine is a web server.

[3] https://iscan.readthedocs.io/

[4] https://spade.glasgow.ac.uk/

[5] http://prosodylab.org/data/bestiary/contour/