Bridging the Gap between Linguists and Technology Developers: Large-Scale, Sociolinguistic Annotation for Dialect and Speaker Recognition*

Christopher Cieri[1], Stephanie Strassel[1], Meghan Glenn[1], Reva Schwartz[2], Wade Shen[3], Joseph Campbell[3]

[1]Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia, PA 19104
{ccieri, strassel, mlglenn}@ldc.upenn.edu

[2]United States Secret Service
Washington, DC
reva.schwartz@usss.dhs.gov

[3]MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421
{swade, jpc}@ll.mit.edu

Abstract

Recent years have seen increased interest within the speaker recognition community in high-level features including, for example, lexical choice, idiomatic expressions or syntactic structures. The promise of speaker recognition in forensic applications drives development toward systems robust to channel differences by selecting features inherently robust to channel difference. Within the language recognition community, there is growing interest in differentiating not only languages but also mutually intelligible dialects of a single language. Decades of research in dialectology suggest that high-level features can enable systems to cluster speakers according to the dialects they speak. The Phanotics (**Ph**onetic **An**notation **o**f **T**ypicality **i**n **C**onversational **S**peech) project seeks to identify high-level features characteristic of American dialects, annotate a corpus for these features, use the data to dialect recognition systems and also use the categorization to create better models for speaker recognition. The data, once published, should be useful to other developers of speaker and dialect recognition systems and to dialectologists and sociolinguists. We expect the methods will generalize well beyond the speakers, dialects, and languages discussed here and should, if successful, provide a model for how linguists and technology developers can collaborate in the future for the benefit of both groups and toward a deeper understanding of how languages vary and change.

1. Introduction

Recent years have seen increased interest within the speaker recognition community in high-level features, so named because they are abstract from the acoustic signal. They include, for example, lexical choice and the presence of idiomatic expressions or syntactic structures. The promise of speaker recognition in forensic applications drives development toward systems robust to channel differences including channel adaptation and the identification of features inherently robust to channel difference. Within the language recognition community, there is growing interest in differentiating not only languages but also mutually intelligible dialects of a single language. Decades of research in dialectology suggest that high-level features can enable systems to cluster speakers according to the dialects they speak. The Phanotics (**Ph**onetic **An**notation **o**f **T**ypicality **i**n **C**onversational **S**peech) project seeks to identify high-level features characteristic of American dialects, annotate a corpus for these features, use the data to dialect recognition systems and also use the categorization to create better models for speaker recognition. Here we report specifically on the project's resource creation efforts.

Phanotics is sponsored by the United States Secret Service with MIT Lincoln Laboratory coordinating the effort and developing the systems. Linguists from Arizona State and Old Dominion universities consult on dialectal phenomena. The Linguistic Data Consortium (LDC) and Appen Pty Ltd. annotated data provided by LDC and others. The effort required to annotate large corpora for many features would be impractical were it not for existing data and annotations and technologies that simplify the annotator's task. The project requires data that have been orthographically transcribed to serve as a guide to potential loci for the features sought. Specifically, we use orthographic transcripts, a pronouncing lexicon and forced-aligner to generate a putative, time-aligned, phonetic transcription of the audio imagining that the speaker's utterances were standard. If the high-level features are described as deviations from standard pronunciation, any locus in which actual pronunciation differs from putative standard is a potential high-level feature. However since complete phonetic transcription is cost-prohibitive, automatic phonetic transcription is not adequately accurate and we lack dialect studies for every difference one might encounter, we instead use these technologies to guide human annotators to expected features.

2. Requirements

This effort requires natural speech from speakers of the target dialects. Initially we focus on distinguishing speakers of African American Vernacular English (AAVE) from other dialects of American English (non-AAVE) though we plan to investigate other American dialects later. We have selected data that was collected to minimize the effect of observation, recordings of subjects engaged in conversations. The project also requires subjects to be categorized according to the dialect spoken. Since the goal of this project is to establish typicality of features by dialect, we must begin with a categorization based on something other than the features themselves. For this we rely on self-reported metadata. Subjects who claim to be native speakers of American English, born and raised in the United States and ethnically African

* This work is sponsored by the Department of Homeland Security under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

American comprise the AAVE pool. The non-AAVE pool contains American English speakers of other ethnicities. We retain the option of removing subjects from either pool if they appear during audit to be miscategorized.

3. Data Selection

The raw audio data come from a number of existing sources. The first is the Mixer corpus (Cieri, et. al. 2006) of conversational telephone speech (CTS) collected at LDC to support the development of speaker recognition technologies robust to changes in language and channel. During recruitment, Mixer subjects provided their age, sex, occupation, cities where born and raised, and ethnicity. Subjects then completed ten or more telephone calls, six minutes in duration, speaking to other subjects, whom they typically did not know, about assigned topics. Bilingual speakers of Arabic, Mandarin, Russian, and Spanish were paired and encouraged to complete some calls in English and some in the other language. Some calls (about 7%) took place in a room where eight or more microphones recorded one side of the conversation. All calls were audited for topic and audio quality but were not generally transcribed. Although Mixer was clearly not designed for the current effort, it is one of the few corpora in which subjects self-report ethnicity. Because the pool contains speakers of multiple American English dialects, who categorized themselves as African American and other ethnicities, it provides a rare opportunity to compare those groups in the same situation. Of course, the data needed to be transcribed, and were, using a specification described below. To date 126 Mixer calls have been transcribed by the Phanotics project. Of these, 35 calls included conversations between two speakers of AAVE while 91 include conversations between one AAVE speaker and one non-AAVE speaker.

The second source of data was the Fisher corpus (Cieri, et. al. 2004) collected at LDC to support the development of speech-to-text technologies within the DARPA EARS program (LDC 2002). During recruitment, Fisher subjects provided their age, sex, native language, and the cities where they were born and raised. During the collection, subjects completed from one to 25 calls, ten minutes in duration, speaking to other participants, whom they typically did not know, about assigned topics. The calls were audited for topic and quality before verbatim, time-aligned orthographic transcripts were produced. Because Fisher was created for another purpose, the development of speech recognition technologies, it lacks crucial information on the ethnicity of the speaker. However, because some of the Fisher subjects were LDC employees, their family, friends, and colleagues, it was possible to identify a handful that could be assigned to an ethnic category after the fact. To date, 171 Fisher calls have been selected for use in this project.

Other potential sources of data include the StoryCorps® Griot (StoryCorps 2008) initiative and several corpora of recording interviews contributed by individual sociolinguists and dialectologists working in communities in the United States. The StoryCorps Griot initiative, funded by the Corporation for Public Broadcasting, is a one-year effort to record interviews of African Americans. This ongoing project will establish nine different recording locations open for up to six weeks each. At each location, African Americans may make appointments to use the recording booth for one hour interviews of friends and family. Potential users receive instructions on how to conduct good interviews and a trained facilitator is present to help the interviews which cover whatever topics the subjects choose. Participants receive a free copy of their interview; other copies are archived and distributed. StoryCorps has agreed to provide selected material to this project in exchange for transcripts. The project has also received data contributions from multiple linguists working in the United States. Project members are reviewing the recordings to determine which are suitable for use in the current effort.

4. Transcription

Since most of the audio data currently lacks transcripts, LDC has designed a transcription specification for this project, similar to the one used for the Fisher collection (LDC 2003). This *Phanotics Quick Transcription* specification emphasizes speed and accuracy.

Annotators begin by segmenting speech, virtually, at the sentence level making sure not to cut too closely to the speech. To improve the results of subsequent forced-alignment, sentences are further segmented if they are longer than eight seconds or contain more than 0.5 seconds of silence internally. Each audio channel, for example each side of a telephone conversation, is segmented independently. Segments may overlap and audio that contains no speech may be left un-segmented. Although these parameters are appropriate for the current task, they constrain the choice of the transcription tool, ruling out for example the very popular Transcriber, which cannot easily handle two channel files and does not permit overlapping segments on a single channel. We use the Xtrans tool (Maeda and Strassel 2004) created at LDC for this effort.

The transcription uses standard orthography, case, and punctuation with a few exceptions. Punctuation is limited to period, question mark, and comma. Double dash marks incomplete sentences and restarts, while single dash marks incomplete words. Proper names are capitalized as are acronyms and strings of letter pronounced separately. Uttered numbers are written as words and not as strings of digits. A limited set of standard contractions are used and non-standard contractions (‘*cause* for *because*), are written as the full word. Obviously mispronounced and idiosyncratic words are tagged with a ‘+’ symbol but, otherwise, no attempt is made to mark dialectal pronunciation since this will be accomplished in the annotation phase. A limited set of non-lexemes, (*um*, *uh*), are used in filled pauses. Speech errors are transcribed exactly as they are produced but the time allocated to transcribe highly diffluent section is limited since these sections will be rejected at the next stage. Annotators are permitted to review disfluent sections at most two times. Background noises are not marked though a limited set of markers is included for speaker noises. Transcribers indicate low confidence in a given section by enclosing it in double parentheses (()).

5. Feature Annotation

The final annotation step for this effort is to identify features that distinguish a given dialect from the standard language. Such features are often described as rules that change the standard form into a non-standard form. The rules apply variably according to internal and external constraints including the lexical identity and morphology of the affected word, its position within a sentence, the phonological environment, the functional effect of the change (for example whether it neutralizes a distinction between two words), the age, sex, socioeconomic class of the speakers, and the dialects they speak. The complete set of rules is too lengthy to discuss here. However, some examples include: the reduction of consonant clusters in final position (*left* → *lef'*, *missed* → *miss*); the deletion of r (*car* → *ca'*), l (*palm* → *pa'm*) and w (*young ones* → *young 'uns*); and, the change of the voiced and voiceless interdental fricatives into voiced and voiceless stops respectively (*bother* → *boda'*).

The preparation described above and customized tools together simplify the annotation process. Since the rules are specified as a→b/x_y ("*a becomes b when preceded by x and followed by y*"), the input and environment parts of the rule (xay) constitute a search term for finding the feature. The input and output parts of the rule (a→b) form a question to be answered: *Did the subject say xay or xby?* This question is generally answered through annotation. Where a phonetic decoder is accurate enough, this question may be answered automatically and we plan to investigate this process in later phases. The tool, SPAAT (Super Phonetic Annotation and Analysis Tool) is designed for rapid annotation and analysis. For each feature under study, it presents the annotator with a list of *regions of interest* (ROI), locations where a rule may have applied, where a feature may be present. Since the transcript and audio have been previously forced-aligned, the annotator can listen to the audio with a small amount of preceding and following context. The annotator's job is to decide whether or not the rule has applied (whether the form is standard or non-standard, whether the feature is present, whether the subject said xay or xby) and to what degree. If the annotator is not certain they may mark ROIs indeterminate. They may also indicate places where problems with noise or voice quality render a decision difficult and places where the tool requests an annotation of a region that is not a true ROI (a false positive).

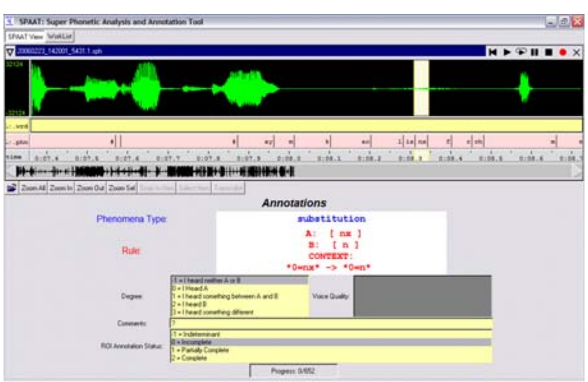

Figure 1: Screen shot of the SPAAT annotation interface.

Figure 1 shows a screen shot of the annotation tool created at MIT Lincoln Laboratory. From top to bottom the panes include a wave form, word level transcription, phone level transcription, time line, and scroll bar overlaid atop a zoomed out wave form. The buttons beneath the scroll bar allow the user to control display and play back by zooming in, zooming out or returning to the view suggested by the forced alignment. The final pane shows the type of phenomena (e.g. deletion, reduction, epenthesis, metathesis, etc.) and the specific rule being annotated at the moment, a series of pick lists for annotator decisions and an input box for comments.

6. Initial Results

The results of our initial pilots have been encouraging. The average time required to annotate an ROI ranges from 15-25 seconds per annotator.

Our approach to measuring interannotator agreement distinguishes *initial agreement* measured at the beginning of an annotation effort to assess the difficulty of a task from agreement measures repeated after thorough documentation has been created and annotators have undergone rigorous training, testing and selection. Since, we have completed our pilots just as this paper is going to print, we can only report briefly on initial agreement. Initial interannotator agreement varies considerably by rule, rule type, annotator and annotator training. Absolute average initial agreement across a pool of five or more annotators, across all rules was 74.49% on a three-way decision where a feature is annotated as present, intermediate or absent. If we convert the decision to two-way (feature is present versus intermediate + absent) initial agreement climbs to 85.54%. Pairwise agreement by chance in three way and two way decisions is, respectively, 11.1% and 25%. Initial two way agreement rates were 83.81% for rules involving substitutions and 91.95% for rules involving reductions and insertions.

Now that we have initial agreement results, the team is expanding the documentation and training program and creating a small gold standard corpus before moving into production annotation. The documentation currently includes a description of each rule, the rule syntax, a prose description with examples and a breakdown of what each possible annotation decision means in the case of that rule. We are currently adding examples with audio to be sure annotators are consistent when they note that a feature is present, intermediate or absent, when an unexpected form is attested, when a decision is too difficult to make and when the region identified is not a true ROI. Once annotators have read this documentation, have trained and have been tested against a gold-standard corpus, we expect significant increases in interannotator agreement. Naturally we retain the option to shift annotators to other tasks if they have difficulty with this task and to exclude rules that apply infrequently, seem to have little distinguishing power or cannot be annotated with adequate confidence.

7. Outcomes

To date the project has: selected approximately 41 hours of audio of conversation telephone speech, transcribed or located transcripts of this audio and completed two pilot phases in which approximately an hour of speech has been annotated for more than a dozen features. Our next goal is to annotate approximately 25 hours of speech from more than 200 speakers for approximately two dozen features to provide data for the first round of system development. Once the data have been used for system development, they will be published by the LDC.

8. Summary

We report on efforts to establish the typicality of some high-level features for dialect and speaker recognition. The project consists of data selection, segmentation, transcription and annotation for the presence of many high-level features that have been shown previously to characterize dialects of American English. The data, once published, should be useful to other developers of speaker and dialect recognition systems and to dialectologists and sociolinguists. We expect the methods will generalize well beyond the speakers, dialects, and languages discussed here and should, if successful, provide a model for how linguists and technology developers can collaborate in the future for the benefit of both groups and toward a deeper understanding of how languages vary and change.

9. References

Christopher Cieri, David Miller, Kevin Walker (2004) The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text, LREC 2004: Fourth International Conference on Language Resources and Evaluation, 2004, Lisbon.

Christopher Cieri, Walt Andrews, Joseph P. Campbell, George Doddington, Jack Godfrey, Shudong Huang, Mark Liberman, Alvin Martin, Hirotaka Nakasone, Mark Przybocki, Kevin Walker (2006). The Mixer and Transcript Reading Corpora: Resources for Multilingual, Crosschannel Speaker Recognition Research, LREC 2006: Fifth International Conference on Language Resources and Evaluation, 2006, Genoa.

LDC (2002) EARS Project Page at the Linguistic Data Consortium, http://projects.ldc.upenn.edu/EARS/

LDC (2003) Rapid Transcription Guidelines, http://projects.ldc.upenn.edu/Transcription/quick-trans/index.html

Kazuaki Maeda and Stephanie Strassel (2004) Annotation Tools for Large-Scale Corpus Development: Using AGTK at the Linguistic Data Consortium, LREC 2004: Fourth International Conference on Language Resources and Evaluation, 2004, Lisbon.

Schwartz, Reva, Wade Shen, Joseph Campbell, Shelley Paget, Julie Vonwiller, Dominique Estival and Christopher Cieri (2007). Construction of a Phonotactic Dialect Corpus using Semiautomatic Annotation, Interspeech. 2007, Antwerp.

StoryCorps (2008) The StoryCorps Griot Web Page, http://www.storycorps.net/special-initiatives/griot