Toward completely automated vowel extraction: Introducing DARLA

Sravana Reddy and James N. Stanford

2015

Abstract

Automatic Speech Recognition (ASR) is reaching further and further into everyday life with Apple's Siri, Google voice search, automated telephone information systems, dictation devices, closed captioning, and other applications. Along with such advances in speech technology, sociolinguists have been considering new methods for alignment and vowel formant extraction, including techniques like the Penn Aligner (Yuan and Liberman, 2008) and the FAVE automated vowel extraction program (Evanini, Isard, and Liberman, 2009, Rosenfelder, Fruehwald, Evanini, and Yuan, 2011). With humans transcribing audio recordings into sentences, these semi-automated methods can produce effective vowel formant measurements (Labov, Rosenfelder, and Fruehwald, 2013). But as the quality of ASR improves, sociolinguistics may be on the brink of another transformative technology: large-scale, completely automated vowel extraction without any need for human transcription. It would then be possible to quickly extract vowels from virtually limitless hours of recordings, such as YouTube, publicly available audio/video archives, and large-scale personal interviews or streaming video. How far away is this transformative moment? In this article, we introduce a fully automated program called DARLA (short for "Dartmouth Linguistic Automation," `http://darla.dartmouth.edu`), which automatically generates transcriptions with ASR and extracts vowels using FAVE. Users simply upload an audio recording of speech, and DARLA produces vowel plots, a table of vowel formants, and probabilities of the phonetic environments for each token. In this paper, we describe DARLA and explore its sociolinguistic applications. We test the system on a dataset of

the US Southern Shift and compare the results with semi-automated methods.

1 Introduction

Existing computational tools for sociophonetics like the Penn Aligner (Yuan and Liberman, 2008) and the ProsodyLab Aligner (Gorman, Howell, and Wagner, 2011) use computational methods to "force align" words and phonemes against speech. Such alignment methods are typically used in a semi-automated manner, i.e., human researchers supply manual transcriptions of sentences, and the aligner maps phonemes to their acoustic representations. More recently, Evanini, Isard, and Liberman (2009) built a program for automated formant measurement, and Rosenfelder, Fruehwald, Evanini, and Yuan (2011) combined it with the Penn Aligner to create a system called FAVE (Forced Alignment & Vowel Extraction). With humans transcribing recordings into sentences, these semi-automated methods can produce effective analyses of field data (Labov, Rosenfelder, and Fruehwald, 2013, Stanford, Severance, and Baclawski, 2014).

Sociolinguistics may now be on the brink of another transformative technology: large-scale, completely automated vowel extraction without any need for human transcription. With such technology, it would be possible to quickly extract vowel formants from virtually limitless hours of recordings, such as YouTube and audio or video archives. Imagine how much more we could learn about dialect variation if we could quickly analyze the millions of hours of audio data around the world in archives and vast publicly available sites. What kinds of new generalizations may be possible? What new kinds of moment-by-moment style shifts might we observe?

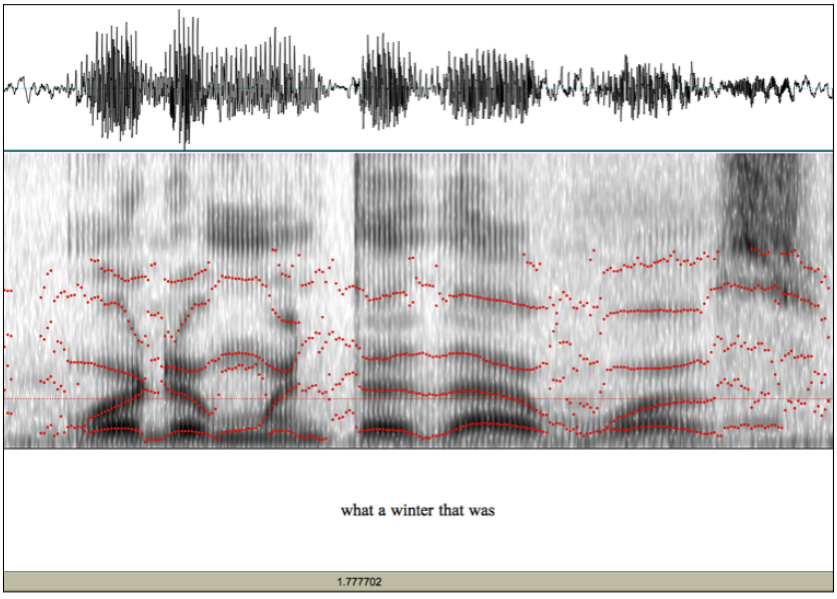
1.1 Prior work

Forced alignment programs take an orthographic transcription and audio as input, and produce a time-alignment of the words and phonemes with the audio. There are several such programs available, including the Penn Aligner, the ProsodyLab Aligner, EasyAlign (Goldman, 2011), and WebMAUS (Kisler, Schiel, and Sloetjes, 2012). All of these systems are built as wrappers around the forced-alignment functions of the HTK speech recognition toolkit (1989-2015), tailored for linguistic

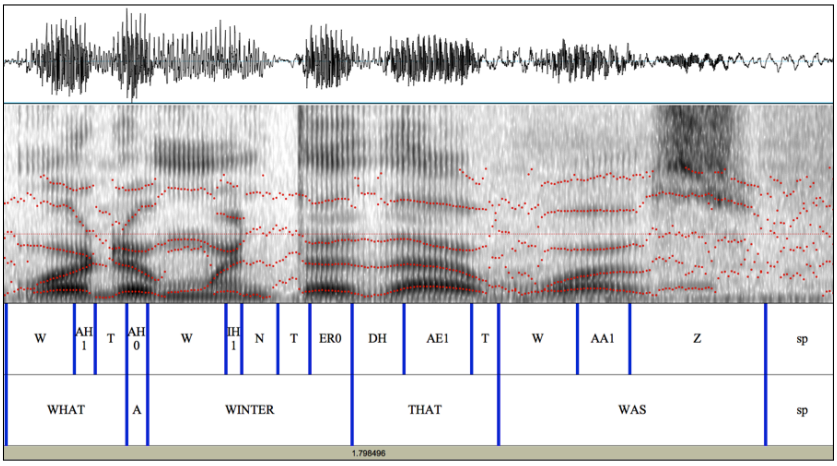
applications.

Methods for obtaining formant measurements have traditionally relied on token-by-token extraction aided by software like Praat (Boersma and Weenink, 2015) that uses Linear Predictive Coding (LPC). With forced alignment tools like the Penn Aligner, a human transcriber produces a sentence-level transcription that is aligned to the audio using acoustic models and a pronunciation dictionary. These alignments help researchers to quickly locate vowels and take formant measurements. University of Pennsylvania researchers have recently developed FAVE, a system that combines the above-described forced alignment (FAVE-Align) with a program, FAVE-Extract, that automatically extracts vowel formant measurements at the appropriate points with LPC. The researcher manually creates a sentence-level transcription, and the semi-automated FAVE system does the rest, returning the final output as a table of vowels and their formant values in the speech recording. An example of this process is shown in Figure 1.

Figure 1: Praat visualization of a semi-automated vowel extraction workflow using FAVE.



(a) Manual transcription to be input to FAVE-Align.



(b) Alignment output from FAVE-Align, which is then input to FAVE-Extract.

vowel	stress	word	F1	F2	F3	time	beg	end	dur
AH	1	WHAT	630.5	1274.5	2012.7	244.61	244.587	244.656	0.069
AE	1	THAT	741.0	1831.4	2908.8	245.24	245.197	245.327	0.130

(c) Final output from FAVE-Extract. The 'stress' column shows whether the vowel has primary, secondary, or no stress in the word, F1, F2, and F3 are the values of the first three formants, and 'time' (in seconds) indicates the point in the audio at which the formants are measured. 'beg', 'end', and 'dur' respectively specify the start and end points, and duration, of the vowel. Not shown: Additional columns with formant measurements at different positions in the vowel, bandwidths, environments, and other information.

Finally, Thomas and Kendall (2007) have developed an R library, `vowels`, and a web interface, NORM, that aids researchers in normalizing and scaling vowel formant measurements as well as producing F1/F2 vowel plots.

Semi-automated alignment and vowel extraction methods therefore represent a significant improvement in efficiency over the largely manual methods that were ubiquitous only a decade ago. No automated or semi-automated method is ever perfect, of course; these methods are prioritizing speed and access to larger data sets, while allowing for some error.

Initial testing suggests that methods like FAVE can produce reasonably accurate vowel measurements of speakers, especially when large amounts of data are processed. Evanini (2009:92) and Evanini et al. (2009) compare manual methods against automated extraction, and show that the amount of error is comparable in most cases to the error found between human analysts. A comparison of such semi-automated methods with Atlas of North American English (Labov, Ash, and Boberg, 2006) data found that the semi-automated results were comparable to manual measurements (Evanini et al., 2009:3-4). In addition, Evanini (2009) points out that the sociolinguistic interpretations (e.g., Northern Cities Shift) are comparable in both approaches. Labov et al. (2013:37-38) compare ANAE formant data against the much larger data set extracted from the Philadelphia Neighborhood Corpus, and find that as the number of tokens increases, the standard error of the mean becomes smaller.

1.2 Research question

Prior work suggests that semi-automated techniques are effective for many types of sociophonetic analyses. Even so, all of these methods still require a significant human workload: manually creating the sentence-level transcriptions.[1] Because of this barrier in time and resources, many large data sets remain unanalyzed. What new knowledge about language variation and change may be waiting in the vast pools of audio recordings that are untranscribed and unanalyzed?

In the present study, we investigate the possibility of removing the human component entirely from the process. Using speech recognition, can vowel extraction be fully automated such that there is no

[1] Transcription is estimated to take 10-15 times the duration of the audio, depending on the quality and type of speech.

need for human intervention at any point? We build a system called DARLA, short for "Dartmouth Linguistic Automation," that automates the entire pipeline from transcribing audio to alignment and formant extraction. The web interface of DARLA is elaborated upon in Reddy and Stanford (2015). In the present paper, we test DARLA on a corpus of audio recordings of northern and southern US speakers and investigate how it compares to semi-automated methods.

1.3 Our idea

The technology of automatic speech recognition is constantly improving, and it is possible that ASR systems will reach human transcription accuracy in a decade or so. However, anyone who has used speech recognition applications such as Apple's Siri, Nuance's Dragon, or automated telephone banking systems knows that the current technology is still very unreliable. These systems are fast and often helpful, but they cannot match the accuracy of human manual transcriptions. In unconstrained natural speech data, errors in speech recognition are quite common.

Despite the imperfection of automatic speech recognition, we believe that sociophoneticians do not need to wait for years to begin taking advantage of the possibilities of large-scale data analyses. The power of ASR can be harnessed right now for some types of sociolinguistic applications. Here's why: whereas applications like dictation software or mobile assistants require that the system captures the words accurately, sociophonetic vowel research generally focuses on a much narrower objective, namely, extracting a representative vowel-space for each speaker, based on stressed vowel tokens. In this paper, we show that current ASR technology is adequate for extracting formant values from stressed vowels for certain sociophonetic problems.

Examples (1-4) show a human transcription compared with ASR output on a subset of the Switchboard corpus (Godfrey and Holliman, 1993). The italicized words highlight the ASR errors. (Details about our speech recognition system, experimental setup and corpus are in Sec. 2.1 and Sec. 3.1.)

(1)
Manual:	give	me	*your*	first	*impressions*
ASR:	give	me	*yours*	first	*impression*

(2)
Manual:	it's	one	of	*those*
ASR:	it's			*close*

(3)

Manual:	no	it's	it's	wood	turning
ASR:	no	*it*	*it*	*would*	*turn it*

(4)

Manual:	and	we	really	don't	spend	on	anything
ASR:	and	we		don't	*depend*	on	anything

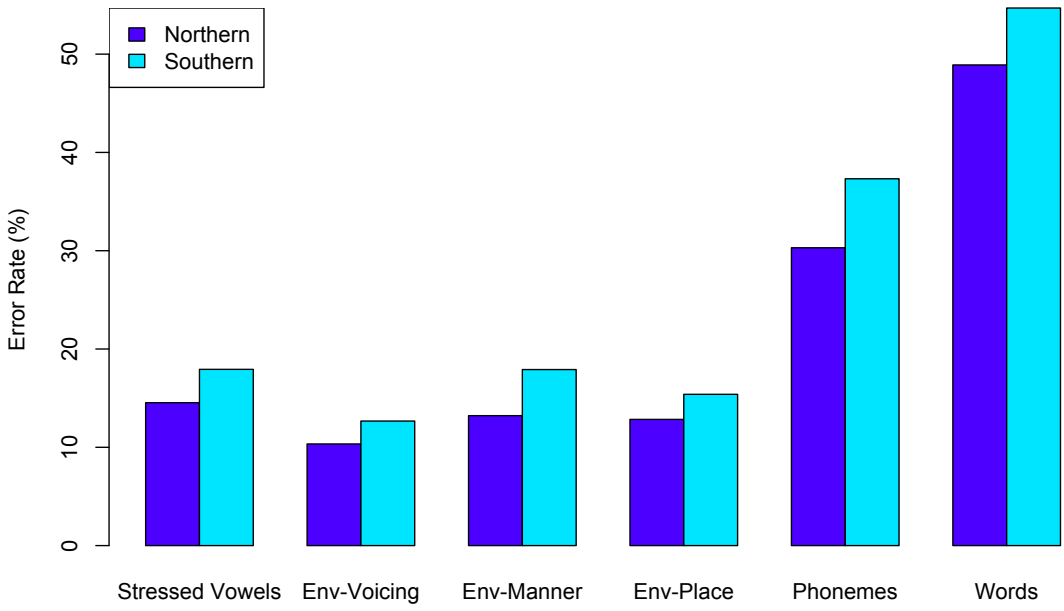
Notice that the ASR errors above do not affect the identity of the vowel being targeted. Rather, the errors affect the identity of the word. The stressed vowels, which are the most common target of sociophonetic analyses, are accurately identified. For example, note the error of 'your' versus 'yours' in example (1). For most practical sociophonetic purposes, it is irrelevant whether the stressed vowel is extracted from the word 'your' or 'yours'. Likewise, in example (3), it is unlikely that any sociophonetic analysis would crucially depend on knowing that the stressed vowel in 'turning' was extracted from 'turning' rather than 'turn it', or that the ASR system outputs the word 'would' instead of 'wood' (example 3). In the last case, the vowel would be accurately measured, although if a researcher chose not to include function words, the vowel token in 'would' would be discarded.

Of course, these examples are not comprehensive: ASR errors that change the identity of the stressed vowel do occur, but to a much lower degree than word errors, and increasing the volume of data can minimize their effects. Figure 2 shows the error rates on the Switchboard corpus used in this study. Notice that the stressed vowel error rate is low in comparison with the error rate at the word or phoneme level.[2] A completely automated vowel extraction process makes it possible to quickly analyze hundreds or thousands of tokens of every vowel in a speaker's vowel space. As a result, errors in a few tokens are likely to become negligible to the overall sociophonetic interpretation of the data for that speaker. In creating our program for automatic vowel extraction, we take advantage of this relatively good stressed vowel accuracy for the purpose of formant analysis.

On the other hand, some vowel environments are important in sociophonetic analyses, such as the presence of sonorant consonants or obstruent+liquid clusters, such as 'close' for 'those' in example (2) above, post-vocalic velar consonants, and so on. Our system accounts for these effects by reporting the probabilities for the phonetic environment for each vowel token, i.e., the likelihood that the segment on either side of the vowel token is in a particular place, manner, or

[2]These error rates are relatively high because the acoustic models are trained on a different corpus – i.e., not Switchboard.

Figure 2: ASR error rates on the Switchboard data (46 Southern and 47 Northern speakers, 55.5 total hours). The error rate for stressed vowels is significantly lower than for words. Errors for phonetic environments of vowels in terms of place, manner, and voicing are also shown.



voicing class. For example, if researchers wish to examine the effects of post-vocalic nasals (e.g., 'pin'/'pen' versus 'pit'/'pet'), they could use the probabilities provided in our system's output to code for these contrasting environments, with similar approaches to examine the effects of post-vocalic velar consonants, pre-vocalic liquids, and so on. Sec. 2.3 describes how the probabilities are computed.

1.4 Completely automated vowel extraction

DARLA's completely automated system takes as input an audio file, and returns data about the vowels, including the formants, phonetic environments, and other information. As such, the manual sentence transcription step (Fig. 1) required for semi-automated systems like FAVE is no longer needed.

We invite other researchers to try the system at `http://darla.dartmouth.edu` and help test it by using their own data sets. Results of such testing will help us to continue to improve DARLA and tailor it to other sociophonetic research questions. Besides research, we believe that DARLA will be useful as a teaching demonstration for introductory linguistics courses and sociolinguistics courses.

2 Methods

DARLA's end-to-end pipeline consists of an ASR engine for transcription of the input audio, forced alignment of the audio with the transcription using the ProsodyLab Aligner, confidence filters and phonetic environment probabilities using classifiers trained on speech data, code built upon FAVE-extract for formant extraction, and vowel space plotting using the R `vowels` package. DARLA is implemented as a user-friendly webpage that allows users to upload their audio files or links to YouTube videos, and receive the results by e-mail. Descriptions of the main components of our system follow.

2.1 Automatic speech recognition engine

DARLA uses a speech recognition system based on Hidden Markov Models (HMMs) (Jelinek, Bahl, and Mercer, 1975), implemented with the latest version of the Carnegie Mellon University (CMU) Sphinx toolkit (2000-2015).

As is standard, the speech signal is represented as a sequence of Mel-Frequency Cepstral Coefficients, or MFCCs (Davis and Mermelstein, 1980). Acoustic frames of 25ms duration are extracted from the audio at overlapping intervals of 10ms, and a set of 12 MFCCs are computed from the Fourier transform of the audio.

The goal of automatic speech recognition is to "decode" the most likely transcription $\hat{w}$ of an input speech recording o, i.e:

$$\hat{w} = \arg\max_{w} P(w|o) = \arg\max_{w} P(w)P(o|w) \tag{1}$$

The quantity $P(w)$ comes from the language model, which is a probability distribution over word sequences in the language. The language model gives a gradient measure of syntactic and semantic "goodness" of a sentence. $P(o|w)$ is given by the acoustic model, which is a distribution that maps phones to their MFCC acoustic representations using HMMs and Gaussian Mixture Models (GMMs). A pronunciation dictionary, which links orthographic spellings of words to their pronunciations, is used in conjunction with the acoustic model. The parameters of these probabilistic models are estimated by training on large speech and text corpora.

We train a triphone acoustic model (which estimates parameters for each phoneme in the context of its left and right neighbors) with a mixture of 32 Gaussians. The choice of acoustic model affects the accuracy of ASR: best results are obtained when the model is trained on data with the same sampling rate and from dialects that are similar to the speech to be transcribed. For this reason, we train separate acoustic models on 8kHz and 16kHz speech, and include the option on the web application to use acoustic models trained on speech from various different dialect regions. The Standard American 8kHz models are trained on the LibriSpeech corpus (Panayotov, Chen, Povey, and Khudanpur, 2015), a collection of US English speech consisting of about 360 hours of audiobooks from the open-access LibriVox project.[3] All training was carried out using SphinxTrain.[4]

The pronunciation dictionary that we use for training and recognition is the CMU Dictionary of Standard American English pronunciations (1993-2015). The language model is trained on the transcripts of the HUB4 broadcast news corpus.

[3] `https://librivox.org`

[4] `https://github.com/cmusphinx/sphinxtrain`

In practice, the language model component in Eq. 1 is typically scaled by a non-negative factor α, the "language weight" (Eq. 2). Larger values of α prioritize the language model and tend to result in more grammatically correct transcriptions, whereas lower values shift the burden to the acoustic model.

$$\hat{w} = \arg\max_{w} P(w|o) = \arg\max_{w} P(w)^{\alpha} P(o|w) \tag{2}$$

The DARLA interface allows users to specify whether their audio consists mainly of free speech and reading passages, for which a larger language weight would be beneficial, or word lists, for which the program uses a lower weight. Pocketsphinx[5] is used for decoding of the speech to produce transcriptions.

2.2 Alignment and token filtering

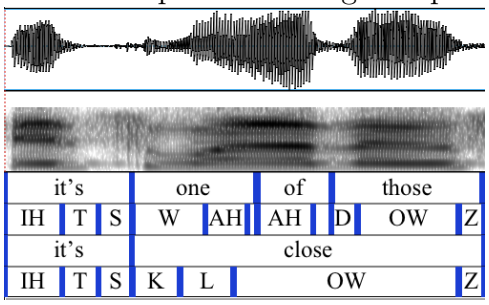
The ProsodyLab aligner with acoustic models trained on a subset (10 hours) of the LibriSpeech corpus is used for the forced-alignment step.

Alignment with noisy transcriptions (such as those produced by ASR) often contain errors even if the vowels are identified correctly. Figure 3 shows an example where the ASR transcribes an input as *it's close* instead of *it's one of those*. Even though the `OW` vowel in the last syllable is correct, the vowel is aligned against multiple words in the speech due to the absence of *one of* in the ASR transcription. Formants extracted from such an alignment will naturally be incorrect. To minimize such errors, we train probabilistic models for each phoneme, based on MFCC features as well as duration, which give the likelihood of any acoustic segment being matched to the hypothesized phoneme. Low-likelihood vowel tokens are removed from the analysis.

Following standard practices in sociophonetics (Baranowski, 2013), DARLA excludes unstressed syllables, grammatical function words and other common lexical items (e.g., *is*, *the*, *I'm*, *and*, etc.), and tokens whose formants have high bandwidths ($>$ 300 Hz) – i.e., tokens which likely have inaccurate formant alignments. The system output includes a file with the unfiltered tokens as well as the filtered measurements.

[5] `https://github.com/cmusphinx/pocketsphinx`

Figure 3: ASR transcription resulting in a poor alignment.



2.3 Phonetic environment probabilities

Sociolinguists are often interested in the phonetic environment of the vowel, such as whether the coda is a velar or the onset is voiced. Researchers in speech recognition have identified methods to classify whether or not any sound frame belongs to one of these distinctive phonetic features (Hasegawa-Johnson, Baker, Borys, Chen, Coogan, Greenberg, Juneja, Kirchhoff, Livescu, Mohan, Muller, Sonmez, and Wang, 2005). Following their work, we train classifiers on each frame in terms of MFCCs, and formant frequencies, amplitudes, and bandwidths for (1) whether the segment is a vowel or consonant and (2) place, manner, and voicing of consonants. The classifiers are logistic regression models trained on the phonetically annotated Switchboard Transcription Project data (Greenberg, Hollenback, and Ellis, 1996) for 8kHz speech, and the TIMIT corpus (Garofolo, Lamel, Fisher, Fiscus, Pallett, Dahlgren, and Zue, 1993) for 16kHz. The probabilities, P(phonetic feature|frame), are averaged over the duration of the segment and reported in DARLA's output.

3 Feasibility test: Fully automated (DARLA) compared to semi-automated (FAVE)

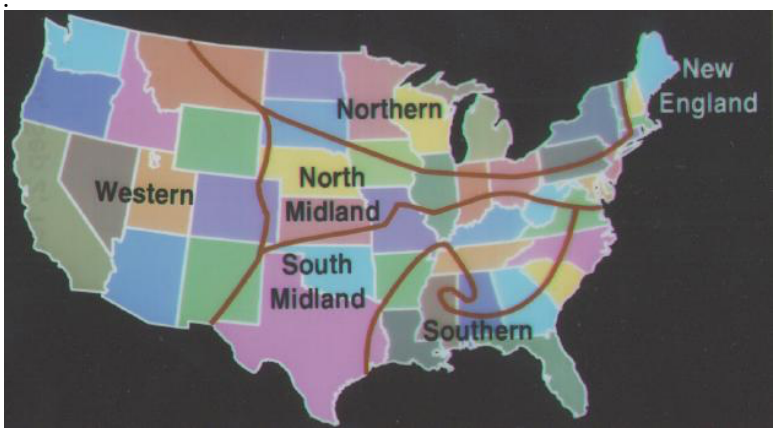
3.1 Experimental Setup

For our feasibility test of DARLA, we examined the U.S. Southern Vowel Shift (SVS) in the Switchboard corpus of phone conversations.

From this corpus, we selected conversations by U.S. Southern and Northern speakers, totaling 46 and 47 speakers respectively. 22 of the Northern speakers and 28 of the Southern speakers are women. This selection of the corpus totals 55 hours and 38 minutes.

We categorized speakers as Southern or Northern according to their home regions in the Switchboard metadata, as compared with the regions identified in the ANAE. Speakers were considered Southern if their Switchboard regions (Fig. 4) were *Southern* or *South Midland*, corresponding approximately to the ANAE categories of *South* and *Texas*.

Figure 4: Dialect regions used in the Switchboard corpus. Reprinted from `http://www.isip.piconepress.com/projects/switchboard/doc/swb_dialects`.



In the present study, we used the 8kHz acoustic models trained on Standard American English speech since we are evaluating the system on a mixed set of dialects.

Fig. 5 shows a sketch of the Southern Shift (Wolfram and Schilling-Estes, 2006, Labov, 1996). In our discussion of the Southern Shift, we use the CMU dictionary's Arpabet vowel notation (Table 1).

For this study of the Southern Shift, we chose to focus on the tense-lax shifts (`EH-EY` and `IH-IY`) and the back vowel fronting of `UW` and `OW`. We hypothesize that the Southern speakers will show greater tense-lax shifts and greater `UW` and `OW` fronting than the Northern speakers. We further hypothesize that both DARLA and FAVE will find evidence of these North-South contrasts. We also note that Kendall and Fridland

Figure 5: The key movements of the U.S. Southern Shift. Adapted from Wolfram and Schilling-Estes (2006:149).

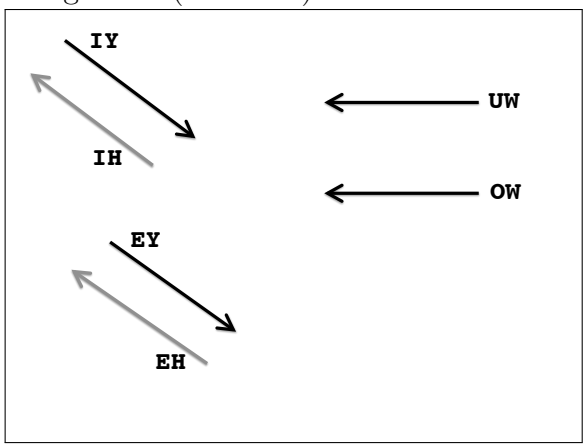


Table 1: Arpabet vowel set.

AA	AE	AH	AO	AW	AY	EH
hot	bat	but	bought	bout	bite	bet
EY	IH	IY	OW	OY	UH	UW
bait	bit	beat	boat	boy	hood	boot

(2012) find that the EH-EY shift is typically more advanced than IH-IY, and therefore expect that EH-EY will be more advanced in this data set. Following Kendall and Fridland, we measure the Euclidean distance (F1/F2 axes) as a measure of the extent of the tense-lax shift.

Our dataset consists of 55 hours and 38 minutes of two-person telephone conversations, with an average of 383 tokens per vowel per speaker. We used DARLA to automatically transcribe and align these Switchboard recordings, and extract F1 and F2 from the stressed vowel tokens after acoustic confidence filtering.

As a control for the feasibility test, we also extracted the formants from the same dataset using the (semi-automated) FAVE system along with manual transcriptions. Specifically, we used FAVE-Align to align the transcriptions provided by Switchboard, and passed those alignments into FAVE-Extract, which provides the formant measurements. We recognize that there are many different approaches to choosing formant extraction points, so we simply used the FAVE default method and applied it consistently to both DARLA and FAVE in this study. Future work could try different measurement points. For both DARLA and FAVE, we normalized the formant data with the Lobanov (speaker-intrinsic) method (Lobanov, 1971, Thomas and Kendall, 2007), and scaled the z-scores to the Hz scale using the constants in NORM.[6]

$$\mathrm{F1}^{\mathrm{scaled}} = 250 + 500 \cdot \frac{\mathrm{F1} - \mathrm{F1}^{\mathrm{min}}}{\mathrm{F1}^{\mathrm{max}} - \mathrm{F1}^{\mathrm{min}}} \tag{3}$$

$$\mathrm{F2}^{\mathrm{scaled}} = 850 + 1400 \cdot \frac{\mathrm{F2} - \mathrm{F2}^{\mathrm{min}}}{\mathrm{F2}^{\mathrm{max}} - \mathrm{F2}^{\mathrm{min}}} \tag{4}$$

3.2 Results

While DARLA showed some transcription errors (see Fig. 2), both DARLA and FAVE generated comparable sociolinguistic analyses of Southern features. This overall result suggests that DARLA can provide usable results for such research questions.

Figure 6 shows that both DARLA and FAVE revealed clear North-South contrasts in EY-EH and IY-IH in the expected directions. For both methods, these shifts appear in the expected Southern Shift directions (EY and IY are lowered and backed, and EH and IH are raised

[6] http://lvc.uoregon.edu/norm/about_norm.php#scaling

and fronted). Likewise, our hypotheses about OW and UW fronting are confirmed as the Southern speakers showed stronger fronting than the Northerners under both systems. The main difference is that DARLA shows slightly weaker Southern effects, as observed in the graphical contrasts between Southern and Northern speakers.

Table 2 quantitatively measures the tense-lax shift by the Euclidean distance between the mean values of the tense and lax vowels in F1/F2 space, averaged over all the speakers in that region. The Southern speakers have a smaller tense-lax distance than the Northern under both DARLA and FAVE, as expected of the Southern Vowel Shift.

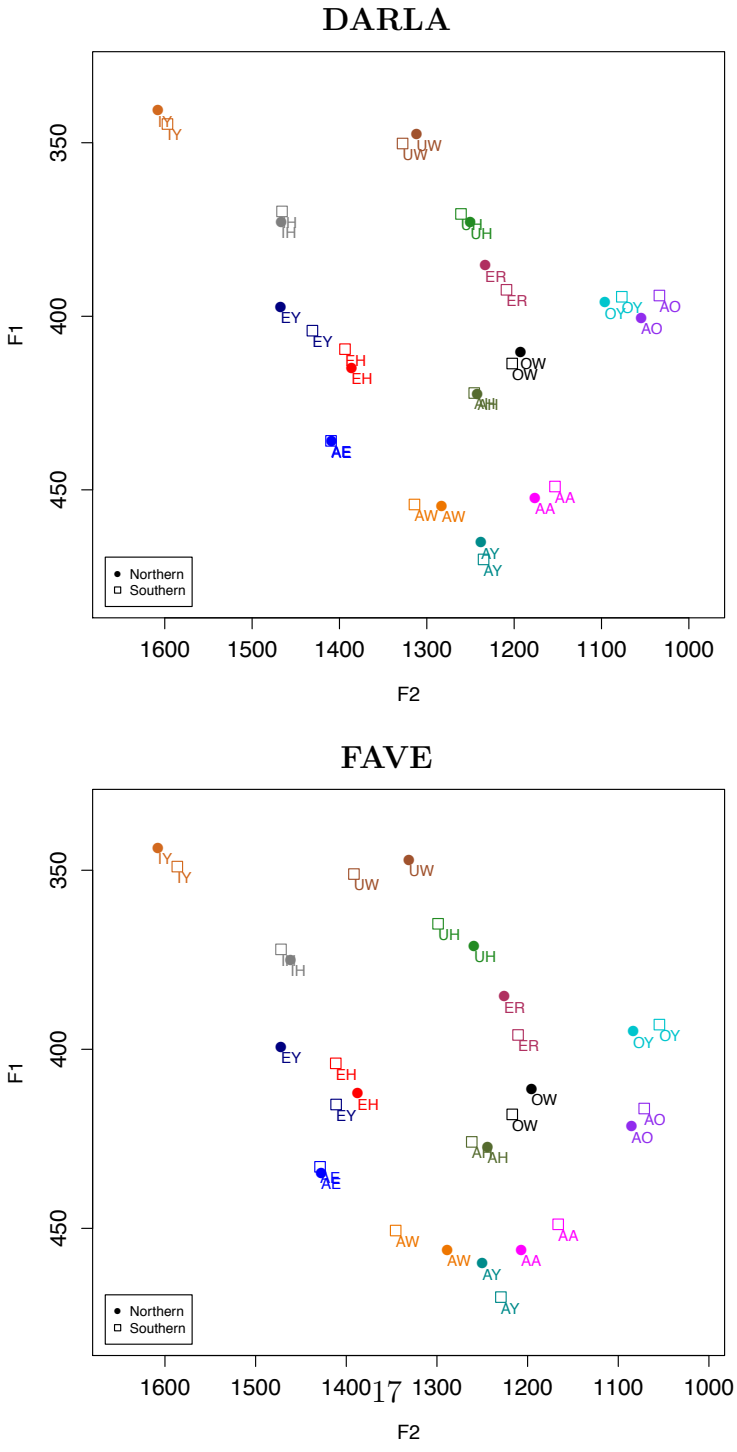
The table also shows the results of the Repeated Measures ANOVA tests of the Northern versus Southern Euclidean distances. The North-South contrast in the EH-EY shift was significant for both methods, but the contrast in IH-IY was only significant in FAVE ($p = 0.011$), not DARLA ($p = 0.284$). For IH-IY, DARLA's results are in the expected direction, but the contrast is not strong enough to be significant. In prior Southern Shift research, EH-EY is typically more advanced than IH-IY (Kendall and Fridland, 2012), and both DARLA and FAVE show this effect as well.

Table 2: Tense-lax shifts.

	FAVE	DARLA
North Mean EH-EY distance	79 Hz	83 Hz
South Mean EH-EY distance	31 Hz	39 Hz
Repeated measures ANOVA	$p = 0.001^{**}$	$p < 0.0001^{***}$
North Mean IH-IY distance	150 Hz	145 Hz
South Mean IH-IY distance	117 Hz	134 Hz
Repeated measures ANOVA	$p = 0.011^{**}$	$p = 0.284$

Figures 7 provides comparisons of the absolute value of the mean differences of vowels using DARLA and FAVE. For many of the vowels, there was not a significant difference in the two measurement techniques. These mean differences were calculated by taking the absolute value of the mean of all measurements of the given vowel by the FAVE methods and comparing it to the mean of all measurements of the same vowel by DARLA. T-tests were also conducted on the set of all FAVE measurements of the given vowel versus the DARLA

Figure 6: DARLA and FAVE formant extraction results. Northern and Southern means are represented by dark circles and unfilled squares respectively.



measurements of that same vowel. Pairwise comparisons were not conducted since the two systems do not always measure the same tokens, as described above. Using pairwise comparisons, prior work on (human) inter-analyst differences has reported the following: Evanini (2009:92-94) finds an average difference of 57.8 HZ for F1 and 126.4 Hz for F2. The other inter-analyst results cited by Evanini are comparable; Labov, Yaeger, and Steiner (1972:32) report a range of 31.5 to 40.5 Hz for F1 and 38 to 84 Hz for F2. Deng, Cui, Pruvenok, Huang, Momen, Chen, and Alwan (2006) find an average inter-analyst difference of 55 Hz for F1 and 69 Hz for F2. Hillenbrand, Getty, Clark, and Wheeler (1995) report inter-analyst differences of 9.2 Hz for F1 and 17.6 Hz for F2. Our mean differences are naturally somewhat lower for two reasons: (a) Figure 7 is based on the mean of the measurement differences, rather than pairwise differences for each measurement, and (b) we Lobanov-normalize the formant values and scale the z-scores to be interpretable in Hz (Equations 3-4).

4 Discussion

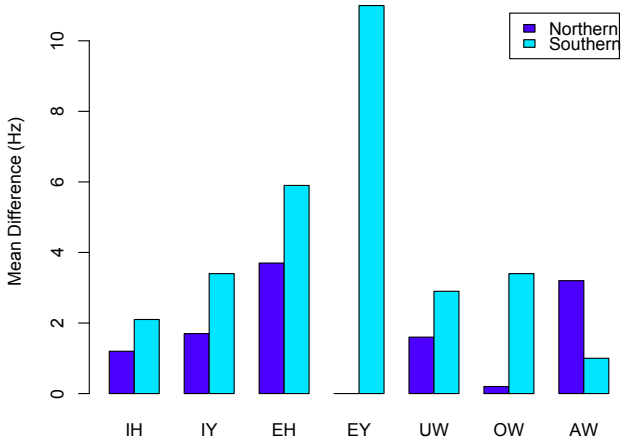
4.1 Bringing replicability and error estimation into sociophonetics

The use of automated methods in sociophonetics suggests the need for a different perspective on measurement error. In traditional manual approaches to sociophonetics, researchers typically treat their extracted formant data as error-free. This assumption is seen in the fact that such research articles rarely include error estimations for the formant extractions. Error estimation usually only appears in the statistical modeling. Yet, testing has shown that data collected by human analysts can be imperfect or inconsistent with other analysts (Sec. 3.2).

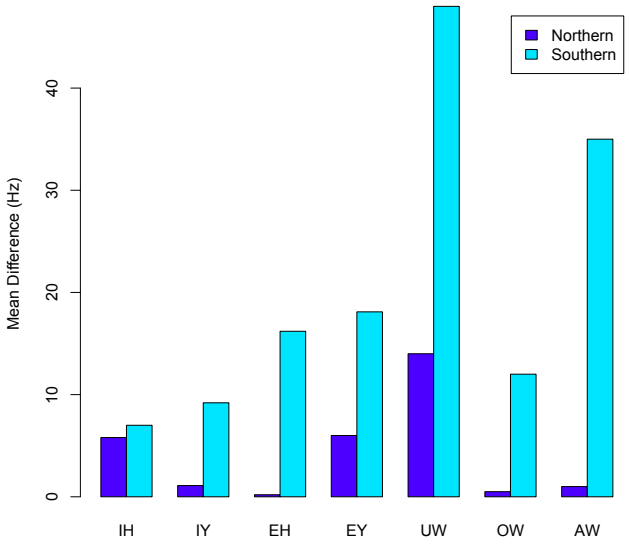
We recognize that some researchers may prefer manual approaches where the human analyst can carefully consider each vowel token, making judgments about which tokens to extract and which extraction point is most appropriate for a particular token. While this approach has produced a great number of valuable analyses and clearly has a number of important strengths, it has limitations in terms of speed and replicability between analysts. It is well known that there is variation in the way different sociophoneticians manually extract vowel

Figure 7: Average differences between vowel formants extracted by DARLA and FAVE. For F1, all the differences except the Northern `EY` and `OW`, and Southern `AW`, are not significant. For F2, the Northern `EH`, `EY`, `OW`, and `AW` differences are significant.

F1



F2



formants, following a range of opinions about extraction points and other variables (Thomas 2011:150-152, Di Paolo, Yaeger-Dror, and Wassink 2011:91, Labov 1994:165).

As Kendall and Fruehwald (2014) point out, the adoption of more automation will make sociophonetics more replicable. This would bring sociophonetics into the realm of fully replicable scientific endeavors, such as physics or chemistry, where measurement error estimations are reported and data sets can be directly compared or checked independently by different research teams. For example, suppose that a sociophonetics researcher manually extracts 20 tokens for each vowel from 50 speakers sampled in a certain population. Suppose that a second researcher works in a nearby population and wants to compare results with the first researcher. If the two researchers had different judgments or philosophies about vowel extraction, the second researcher might have to measure formants at many different extraction points, through trial and error, from the first researcher's data set in order to replicate the results. Since such a task would be arduous, sociophoneticians often rely on qualitative comparisons of other researchers' published work, rather than directly comparing the data sets. Therefore, while fully automated vowel extraction necessarily brings in some sources of error, it has the advantage of allowing for quick, consistent, and complete replication across data sets as long as the ASR, alignment, and formant extraction parameters are reported.

4.2 Future directions

We are currently exploring recent advances in ASR that use deep neural networks (Hinton, Deng, Yu, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath, Dahl, and Kingsbury, 2012) for acoustic modeling. Such methods have demonstrated remarkable improvements in recognition accuracy over the traditional HMM+GMM models, and will likely improve the overall performance of our system. Eventually, we would like to expand the scope of DARLA beyond English, and implement automated methods for analyzing sociophonetic variables besides vowels, drawing upon previous research in measuring rhoticity (Hesselwood, Plug, and Tickle, 2010), voice onset time (Sonderegger and Keshet, 2012), glottal source (Kane, 2012), and other features.

5 Conclusion

This project has taken a first step toward completely automatic vowel formant extraction, introducing a system called DARLA (Dartmouth Linguistic Automation). Our results suggest that completely automated ASR systems like DARLA can be used for extracting formant data to answer meaningful sociolinguistic questions. ASR technology is not reliable enough for consistently accurate transcriptions, but it is reliable enough for certain types of research. In particular, we note that sociophonetic vowel research is usually focused on finding representative vowel spaces from stressed vowels. For such goals, perfect accuracy in transcriptions is not usually necessary. Obtaining representative vowel tokens may often be sufficient, as long as the likely phonetic environment for each token is included along with probabilities.

In this study, we conducted a feasibility test of DARLA using U.S. Southern and Northern speakers from the Switchboard telephone corpus. As a control, we extracted formants from the same dataset using the semi-automatic FAVE system, which requires a human analyst to manually create sentence-level transcriptions. Both DARLA and FAVE found clear evidence of the Southern Vowel Shift in the Southern speakers, although the shifts were somewhat weaker in DARLA.

No automated system is perfect, but we believe that the potential access to large-scale datasets makes these types of automated approaches worthwhile, even though they differ from traditional manual approaches. Many other scientific fields like physics and astronomy regularly report error estimates in their measurements, not just in their statistical modeling. Such error estimates and accessible automation techniques could make it possible to analyze vast amounts of audio data in fully replicable research. With continuing advances in ASR technology in coming years, the widespread use of completely automated vowel extraction for sociophonetic research may not be far away.

Acknowledgements

We are grateful to the anonymous reviewers for their suggestions, and to the various users of DARLA for feedback. Irene Feng assisted in building the web interface. The first author was supported by a Neukom Fellowship at Dartmouth, and development of DARLA is be-

ing sponsored by a Neukom CompX grant. The computing cluster used for training the ASR models and running experiments was made available by NSF award CNS-1205521.

References

Baranowski, M. (2013): "Sociophonetics," in R. Bayley, R. Cameron, and C. Lucas, eds., *The Oxford Handbook of Sociolinguistics*, Oxford: Oxford University Press, 403–424.

Boersma, P. and D. Weenink (2015): "Praat: doing phonetics by computer [computer program]," available at `http://www.praat.org`.

Cambridge University (1989-2015): "HTK Hidden Markov Model Toolkit," available from `http://htk.eng.cam.ac.uk`.

Carnegie Mellon University (1993-2015): "CMU Pronouncing Dictionary," available from `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

Carnegie Mellon University (2000-2015): "CMU Sphinx Speech Recognition Toolkit," available from `http://cmusphinx.sourceforge.net`.

Davis, S. B. and P. Mermelstein (1980): "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28, 357–366.

Deng, L., X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan (2006): "A database of vocal tract resonance trajectories for research in speech processing," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, `http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1660034`.

Di Paolo, M., M. Yaeger-Dror, and A. B. Wassink (2011): "Analyzing vowels," in M. Di Paolo and M. Yaeger-Dror, eds., *Sociophonetics: A student's guide*, London: Routledge.

Evanini, K. (2009): *The permeability of dialect boundaries: A case study of the region surrounding Erie, Pennsylvania*, Ph.D. thesis, University of Pennsylvania.

Evanini, K., S. Isard, and M. Liberman (2009): "Automatic formant extraction for sociolinguistic analysis of large corpora," in *Proceedings of Interspeech*, http://www.isca-speech.org/archive/interspeech_2009/i09_1655.html.

Garofolo, J., L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue (1993): *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*, Philadelphia: Linguistic Data Consortium.

Godfrey, J. and E. Holliman (1993): *Switchboard-1 Release 2 LDC97S62*, Philadelphia: Linguistic Data Consortium.

Goldman, J.-P. (2011): "Easyalign: an automatic phonetic alignment tool under Praat," in *Proceedings of Interspeech*, http://www.isca-speech.org/archive/interspeech_2011/i11_3233.html.

Gorman, K., J. Howell, and M. Wagner (2011): "Prosodylab-Aligner: a tool for forced alignment of laboratory speech," *Canadian Acoustics*, 39, 192–193.

Greenberg, S., J. Hollenback, and D. Ellis (1996): "Insights into spoken language gleaned from phonetic transcriptions of the Switchboard corpus," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, http://www.silicon-speech.com/Media/PDF/1996_Greenberg%26CO_InsightsSpokenLanguage.pdf.

Hasegawa-Johnson, M., J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez, and T. Wang (2005): "Landmark-based speech recognition: Report of the 2004 Johns Hopkins Summer Workshop," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1415088.

Hesselwood, B., L. Plug, and A. Tickle (2010): "Assessing rhoticity using auditory, acoustic and psychoacoustic methods," in *Proceedings*

of Methods XIII: Papers from the 13th International Conference on Methods in Dialectology.

Hillenbrand, J., L. Getty, M. Clark, and K. Wheeler (1995): "Acoustic characteristics of American English vowels," *Journal of the Acoustical Society of America*, 97, 3099–3111.

Hinton, G., L. Deng, D. Yu, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, G. Dahl, and B. Kingsbury (2012): "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, 29, 82–97.

Jelinek, F., L. Bahl, and R. Mercer (1975): "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Transactions on Information Theory*, 21, 250–256.

Kane, J. (2012): *Tools for analysing the voice: Developments in glottal source and voice quality analysis*, Ph.D. thesis, Trinity College Dublin.

Kendall, T. and V. Fridland (2012): "Variation in perception and production of mid front vowels in the U.S. Southern Vowel Shift," *Journal of Phonetics*, 40, 289–306.

Kendall, T. and J. Fruehwald (2014): "Towards best practices in sociophonetics (with Marianna Di Paolo)," in *New Ways of Analyzing Variation (NWAV) 43*, Chicago.

Kisler, T., F. Schiel, and H. Sloetjes (2012): "Signal processing via web services: the use case WebMAUS," in *Digital Humanities Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts*, `http://www.clarin-d.de/images/workshops/proceedingssoasforthehumanities.pdf`.

Labov, W. (1994): *Principles of linguistic change. Volume 1: Internal factors*, Oxford: Blackwell.

Labov, W. (1996): "The organization of dialect diversity in North America," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, `http://www.ling.upenn.edu/phono_atlas/ICSLP4.html`.

Labov, W., S. Ash, and C. Boberg (2006): *The Atlas of North American English (ANAE)*, Berlin: Mouton.

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

Labov, W., M. Yaeger, and R. Steiner (1972): "A quantitative study of sound change in progress," Report on NSF Contract NSF-GS-3287.

Lobanov, B. M. (1971): "Classification of Russian vowels spoken by different speakers," *Journal of the Acoustical Society of America*, 49, 606–608.

Panayotov, V., G. Chen, D. Povey, and S. Khudanpur (2015): "LibriSpeech: an ASR corpus based on public domain audio books," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Reddy, S. and J. N. Stanford (2015): "A web application for automated dialect analysis," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) - Demos*, `http://aclweb.org/anthology/N/N15/#3000`.

Rosenfelder, I., J. Fruehwald, K. Evanini, and J. Yuan (2011): "FAVE (Forced Alignment and Vowel Extraction) Program Suite," available at `http://fave.ling.upenn.edu`.

Sonderegger, M. and J. Keshet (2012): "Automatic measurement of voice onset time using discriminative structured prediction," *Journal of the Acoustical Society of America*, 132, 3965–3979.

Stanford, J., N. Severance, and K. Baclawski (2014): "Multiple vectors of unidirectional dialect change in eastern New England." *Language Variation and Change*, 26, 103–140.

Thomas, E. (2011): *Sociophonetics: An introduction*, New York: Palgrave Macmillan.

Thomas, E. and T. Kendall (2007): "NORM: The vowel normalization and plotting suite [online resource]," available at `http://ncslaap.lib.ncsu.edu/tools/norm`.

Wolfram, W. and N. Schilling-Estes (2006): *American English (2nd edition)*, Malden, MA: Blackwell.

Yuan, J. and M. Liberman (2008): "Speaker identification on the SCOTUS corpus," *Journal of the Acoustical Society of America*, 123, 3878.