

Harvesting Speech Datasets for Linguistic Research on the Web

White Paper

Mats Rooth, Jonathan Howell and Michael Wagner

1. Introduction

Already a huge amount of human speech is available on the web, in the form of podcasts, radio and television content, university lectures, political speeches and much else. If we could somehow observe it, there is the potential of confronting social-science theories of spoken language with data on an unprecedented scale. Our particular interest is in the pervasive and subtle phenomenon of *prosody* (rhythm, stress and intonation). Theories about prosody ultimately capture correlations between acoustic form and grammar (phonology, morphology, syntax, semantics, and pragmatics). We would like to test and refine such theories using large web-sourced datasets---in particular, datasets consisting of hundreds or thousands of utterances of a single short word sequence.

We leveraged our work off web sites that index audio content with transcriptions obtained with automatic speech recognition (ASR). By providing textual transcriptions, content providers intend to make spoken language searchable, so that users can find content they are interested in. While the transcriptions that are provided by current technology have variable quality at the sentence level, accuracy is often better than 50% at the level of short, common word sequences. This makes it possible to create targeted datasets from web sources, using a combination of simple programs and hand work.

The purpose of collecting the targeted datasets is to evaluate hypothesized correlations between acoustic form and grammatical and contextual features, and to identify the particular acoustic features (such as pitch, duration, intensity, or vowel quality) that are significant in marking prosodic distinctions. To do this, we use a machine learning classification paradigm, where a classifier is trained to make a binary distinction based on acoustic measures. Trying to create an acoustic classifier that correlates with a grammatical/contextual feature provides a powerful test of the significance of the feature in the output side of the linguistic system.

This document is organized as follows. Section 2 describes the “harvest” portion of our workflow, where a preliminary dataset is collected from web sources, mainly by automatic methods. Section 3 is concerned with phonetic analysis, acoustic measurements, and machine

learning. Section 4 addresses the question of comparing results that are obtained from web-derived datasets with results from data obtained in the lab. Section 5 (not available in this draft) will describe the project organization, lessons learned, prospects for the web-dataset methodology, and the like.

2. Harvest procedure

Using Unix tools, we implemented a harvest procedure that automatically collects a preliminary dataset for a fixed target. Each datum is a 30-second audio snippet that surrounds a possible token of the target word sequence. The procedure interacts with websites that index audio content transcriptions derived by ASR, and in its outline mimics what a human user would do to retrieve audio content from the website. Figure 1 gives part of a page that results from searching for the word sequence “in my opinion” at audio.weei.com, a web site for the Boston sports radio station WEEI. This page represents then hits---for each hit the name of a radio program is displayed, together with the time offset for the target in the audio file. Figure 2 gives a similar page at mediasearch.wnyc.org, a site of the New York public radio station WNYC.

A user clicking on the first hit in Figure 1 brings up the page in Figure 3, which is a flash-based audio player, together with a transcription of a context that surrounds the target. The user can click to play the target or the whole file, and is invited to distribute a link on social media, or download the audio file.

A user who wanted to collect tokens of the word sequence “in my opinion” on WEEI or WNYC would visit each of about 50 pages that display ten hits each, and visit pages for individual hits by following links from these pages. At each individual page, an mp3 audio file would be downloaded. Time offsets and other data such as the surrounding context in the ASR transcription would also be recorded. In mimicking these steps, our harvest procedure retrieves web pages using *curl*, a command line program that retrieves pages designated by URL. Simple text processing is used to extract information such as the time offset and the URL of the mp3 from the html-encoded pages.

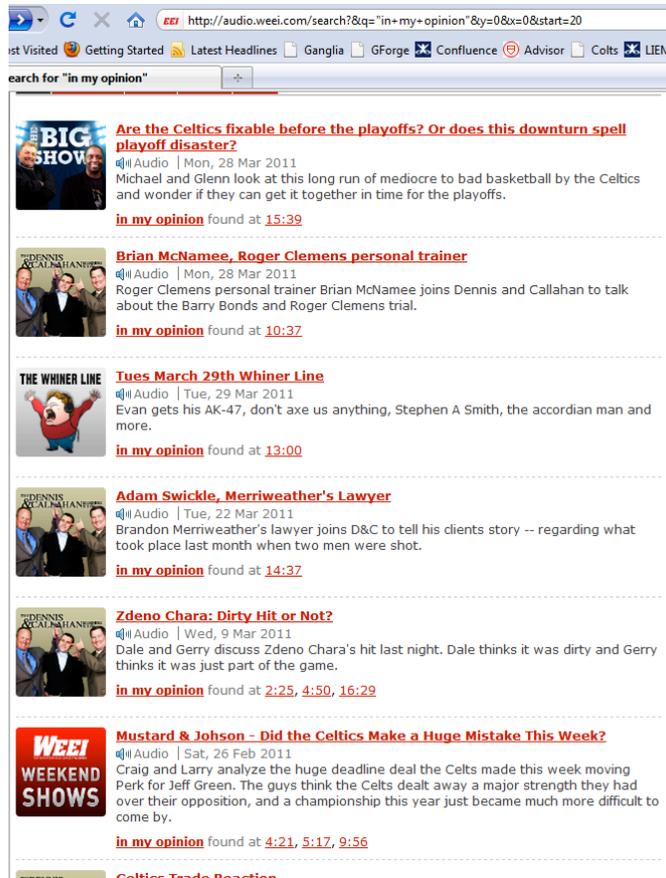


Figure 1. Part of a browser display at audio.veei.com with hits for “in my opinion”. A time offset is included for each hit. The url encodes the target as “in+my+opinion”, while “start=20” requests hits 20 through 29.

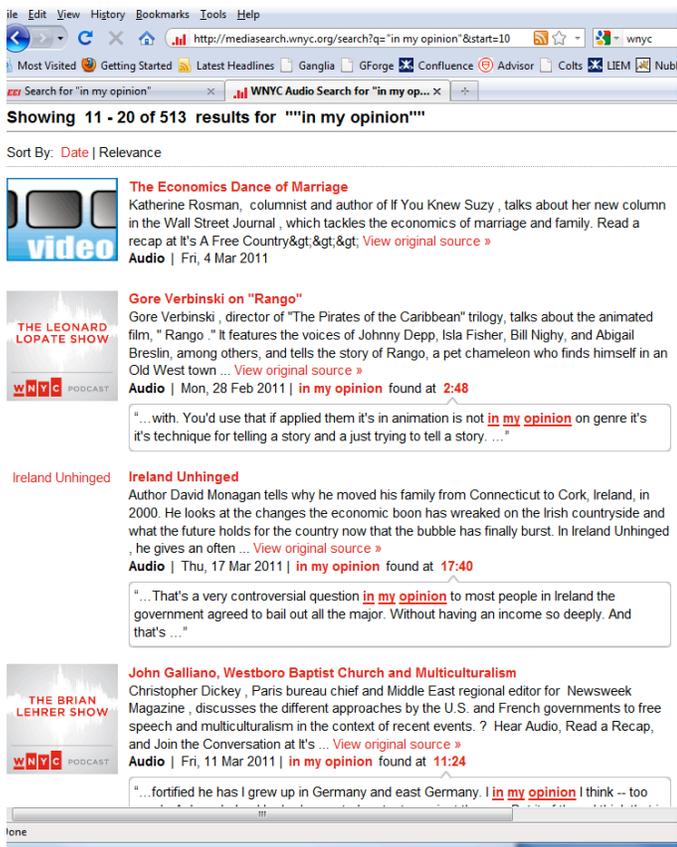


Figure 2. Page at mediasearch.wnyc.org displaying hits for “in my opinion”. Time offsets and an ASR transcription of the context are included.

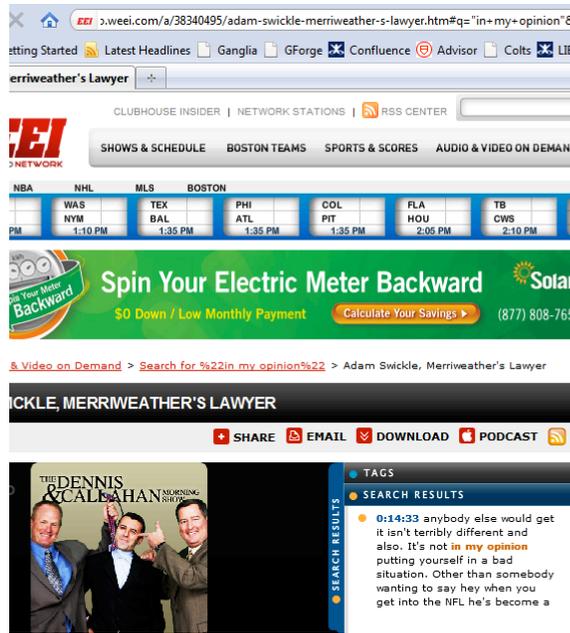


Figure 3. A player at audio.ween.com displaying a hit for “in my opinion”. The SHARE button creates links in social media. DOWNLOAD enables the user to retrieve an mp3 audio file. The SEARCH RESULTS box gives the time offset and an ASR transcription. Clicking on an individual word starts the player shortly before the word.

We exemplify the automatic harvest procedure with a retrieval of about 450 possible tokens of the word sequence “and I think” at mediasearch.wnyc.org. Hits are designated by natural numbers, and in file names these indices are appended to a base name derived from the target. Thus *andithink466.param* is a file associated with hit 466. As shown in Figure 5, this file records the URL of the web page for the hit, the URL of the mp3, the time location of the hit in the audio, and the left and right contexts for the token. The file *andithink460.hits* is an html file representing ten hits, including hit 466. Figure 6 displays the commands that were executed in retrieving *andithink460.hits*. The page is retrieved with *curl* using the target URL as a parameter. After retrieving a file, the procedure sleeps for 25 seconds, to control the rate at which server is accessed.

```

INDEX      466
HIT        http://mediasearch.wnyc.org/m/37248530/world-on-the-edge.htm
MP3        http://feeds.wnyc.org/~r/wnyc_b1/~5/ExTC2araG6I/b1022511cpod.mp3
SEEK       758.939
TS         12:44
LEFTCONTEXT      you know we were climate took matters. We are changing our
                  position
RIGHTCONTEXT     that was one of the most dramatic things to come out of --
                  that's extraordinary day heat wave and drought that Russia

```

Figure 5. The file *andithink466.param*. The values of HIT and MP3 are urls. SEEK and TS specify the time offset in different notations. LEFTCONTEXT and RIGHTCONTEXT are the context derived from speech recognition that is specified on the WNYC page. The context is nearly correct in this case.

```

echo "getting Data/andIthink460.hits with curl" >> Log/andIthink1.loga
curl --verbose --location --output Data/andIthink460.hits
  'http://mediasearch.wnyc.org/search?q=%22and%20I%20think%22&start=460'
sleep 25
cat Data/andIthink460.hits | awk -f extracthitpages.awk BASE="andIthink"
  INDEX=460 >> andIthink2.sh

```

Figure 6. Commands that retrieve *andithink460.hits*, an html page representing ten hits for “and I think”, and extract parameters from that page. Text processing is performed with the awk programming language.

```

echo "getting Data/andIthink460.hits with curl" >> Log/andIthink1.loga
curl --verbose --location --output Data/andIthink460.hits
  'http://mediasearch.wnyc.org/search?q=%22and%20I%20think%22&start=460'
sleep 25
cat Data/andIthink460.hits | awk -f extracthitpages.awk BASE="andIthink"
  INDEX=460 >> andIthink2.sh

```

Figure 7. Commands from *andithink1.sh* that retrieve *andithink460.hits* (which is an html page representing ten hits for “and I think”) and extract parameters from that page.

The lines in Figure 6 are a part of a bash shell program *andithink1.sh* that retrieves fifty html files *andithink0.hits* ... *andithink490.hits*, each of them representing 50 hits. The shell program is created with an awk program *master.awk*, parameterized as follows:

```
gawk -f master.awk -v TARGET='and+I+think' -v RESULTS=500
```

The TARGET parameter sets the target word sequence, while the RESULTS parameter indicates the number of hits to be retrieved. The result of this call is the shell program *andithink1.sh*, as well as an additional program *andithink.sh* that is used to trigger the entire retrieval.

The last line in Figure 6 passes the hits file *andithink460.hits* through an awk program *extracthitpages.awk*, resulting in a sequence of shell commands that are partially shown in Figure 7. The line calling *curl* retrieves the html page for hit 461, putting it in *andithink461.hit*. The remaining lines write information into a tab-delimited parameter file *andithink461.param*. HIT is the URL of the hit page; SEEK and TS are time offsets in different notations; LEFTCONTEXT and RIGHTCONTEXT give the context for the target that is shown for hit 461 on *andithink460.hits*. MP3 is the url of the mp3 audio file, which is extracted from *andithink461.hit* by an awk program *extractmp3name.awk*. The other information is extracted from *andithink460.hits* by *extracthitpages.awk*. The shell program *andithink2.sh* contains a sequence of commands like this for each hit index.

```
echo "INDEX      461" >> Data/andithink461.param
echo "HIT        http://mediasearch.wnyc.org/m/36985155/analysis-of-events-in-egypt.htm" >> Data/andithink461.param
curl --location --output Data/andithink461.hit
      http://mediasearch.wnyc.org/m/36985155/analysis-of-events-in-egypt.htm
sleep 25
cat Data/andithink461.hit | awk -f extractmp3name.awk >>
      Data/andithink461.param

echo "SEEK       74.359" >> Data/andithink461.param
echo "TS         1:18" >> Data/andithink461.param
echo "LEFTCONTEXT      yet we've seen that demonstrated. They're trying to
      stake out some ground" >> Data/andithink461.param
echo "RIGHTCONTEXT     that's what these statements mean at this point. On
      where they convey that they understand egyptians don't trust us to move our
      " >> Data/andithink461.param
```

Figure 8. Commands in shell program *andithink2.sh* pertaining to hit 461.

The final steps are retrieving an mp3 from the server, and cutting the mp3 to a shorter 30-second segment that surrounds the target. The mp3 is retrieved with *curl*, while cutting is accomplished by *cutmp3*, a command line program that manipulates mp3 files. Both of these steps are triggered by the shell program *andithink.sh*, which also calls *andithink1.sh* and *andithink2.sh*.

count	description
50	html files <i>andithink0.hits</i> ... <i>andithink50.hits</i> with 10 hits each
460	html hit files <i>andithink1.hit</i> ... <i>andithink500.hit</i>
462	param files <i>andithink1.param</i> ... <i>andithink500.param</i>
456	mp3 files <i>andithink1.mp3</i> ... <i>andithink500.mp3</i>
450	cut mp3 files <i>andithink1-b.mp3</i> ... <i>andithink500-b.mp3</i>
3	shell programs <i>andithink.sh</i> <i>andithink1.sh</i> <i>andithink2.sh</i>
?	log files ...

Figure 9. Files created in a harvest for the target “and I think”.

Figure 9 tabulates the files that were created in the *andithink* harvest. The sum size of the files is 4.2G. The time duration for the harvest is about eleven hours. Much of this time is occupied by the process sleeping. The entire process it triggered with the call to *master.awk* given above, followed by a call to *andithink.sh*. Thus for the user, collecting the information in Figure 8 is accomplished easily.

In summary, the harvest component uses command line programs to retrieve web pages and audio files and to cut audio files, uses the awk programming language for text processing, and uses a make, awk, and bash scripts to control the process. We found these simple methods to be optimal because of their flexibility.

The harvest component was developed incrementally at Cornell over the course of a year. It is being used regularly for harvests at Cornell and at McGill, and has been used in a teaching context in a seminar at Cornell. Using the component requires no more than familiarity with the Unix shell environment. The procedure has been confirmed run in Redhat Enterprise Linux, Mac/OS X, and Windows/Cygwin operating system environments.

Because different websites represent information in slightly different ways, the text processing programs *extracthitpages.awk* and *extractmp3name.awk* need to be rewritten for each site, and may also need to be debugged when a site changes its representation. Sampling data from different sources is clearly advantageous, because it increases the diversity and size of the dataset.

In the final version of this white paper, we plan to include information about multiple working specializations to different websites.

The text processing programs use line-based text processing, and do not parse the html. We experimented with software systems that parse the html, and with *xgawk*, a version of *awk* that iterates through html elements, rather than lines in the file. With the former, there were difficulties with apparent ill-formedness in the html. We had success with *xgawk*, but opted for the simpler solution because of difficulty in getting it to run consistently on all of the operating systems we work with.

In the time remaining until the June 2011, we anticipate not making substantial additions to the software, except providing specializations to multiple sites. There is time for substantial additional software development in the NSF-funded project at Cornell, which was funded with a two-year period ending in July 2012. Here we will sketch plans for further software development.

Automated filtering and segmentation for fixed targets

Suppose a hundred tokens for a fixed target (such as “than I did” or “in my opinion”) have been hand labeled with word and phoneme boundaries, and that many more possible tokens are available from the web harvest. The hand-labeled data provide a training set that should be useful in reducing the amount of human labor required in extending the dataset. We would like to automatically separate actual tokens from non-tokens, identify the time interval occupied by a target token, and perform phone-level segmentation of the target.

In a pilot investigation, we trained an HMM recognizer on 90 tokens from the *thanididl* dataset using the HTK toolkit, and used the recognizer to obtain a phone-level segmentation of additional tokens. While the results have not been evaluated, HMM models trained in this way on the phone sequence in the fixed target are impressionistically appear good at segmenting novel tokens. We plan to validate the automatic segmentation by determining whether results in SVM classification focus are as good with automatic segmentation as with hand segmentation. Second, with a probabilistic acoustic model of the target available, it is possible to evaluate information-theoretically the hypothesis that a given segment of the signal contains a token of the target. This is a possible basis for an automatic filtering algorithm. Similarly, it should be possible to use the HMM acoustic model to find the interval corresponding to the target in the thirty-second sound snippet.

Deduping

Online radio content often includes multiple literal copies of the same audio segment. Currently, we are identifying duplicates in an ad-hoc way, either in the course of transcription by a human annotator, or after transcription. For an automatic procedure, two sources of information are available. Speech recognition transcriptions, though imperfect, often are identical or similar for audio segments that are actual copies. Since the ASR-determined context is being recovered from the website, it is possible to compare the contexts for different tokens by a measure such as string-edit distance, to identify possible duplicates. Second, one can try to compare the actual speech signals, by searching for intervals that are copies. This may be complicated by transformations in the signal, including noise, digital encoding and decoding, and perhaps time compression.

An interesting intermediate case is provided by distinct utterances of the same sentences, either by different announcers or the same one. One would like to identify these and mark them in the dataset, other than eliminate them.

We call the process of eliminating literal duplicates in the dataset *deduping*. Currently, we are beginning work on it by collecting examples of duplicate tokens.

Alignment of complete transcripts

Some websites provide complete ASR transcripts for programs. While the transcripts are imperfect, they contain sub-intervals that agree with the actual words uttered, and even the incorrect parts should agree in an approximate way with the signal, because they are obtained with speech recognition. Deriving an alignment from a transcription, a signal, and a pronouncing dictionary is standard methodology in training a speech recognizer. We would like to experiment with aligning the entire transcript, as an alternative way finding the time limits of the target in the signal.

- c. I did more **this time** [than I did] **last time** Non-subject focus (category “ns”)

In other words, we have a straightforward way of classifying the data into semantic categories independent of prosodic information. Previous prominence predictors, which attempt to compute the location(s) of prominence mainly from the text, use various properties, from a word’s part of speech (Altenberg 1987, Hirschberg 1993, Conkie et al. 1999) or predictability (Pan & McKeown 1999, Pan & Hirschberg 2001, Gregory & Altun 2004, Brenier 2008) to syntactic embeddedness (Chen & Hasegawa-Johnson 2004) or position in a sentence (Sun 2002, Gregory & Altun 2004). The single criterion used here—subject co-reference—is straightforwardly computed and is well understood in contemporary linguistic theory.

Prominence detectors, which attempt to compute prominence from acoustic information, typically concentrate on measures of fundamental frequency, the physical correlate of pitch, as do the majority of phonetic studies of prominence. Beginning with Fry (1955, 1958) and later by Lieberman (1960), it became clear that other measures, including amplitude and duration, provide at least partial cues as well. More and more acoustic measures have appeared in the literature on acoustic prominence, including measures of voice quality (“spectral tilt”; Sluijter and van Heuven 1996a,b, 1997; Campbell & Beckman 1997; Heldner 2001, 2003; Mo 2008), vocal tract resonances of particular vowels (Kim & Cole 2005; Erikson 2002; Cho 2005; Cole *et al.* 2007; Mo 2008, 2009) and stop closure duration (Cole *et al.* 2007). “Hyperarticulation” and “featural enhancement” theories (e.g. de Jong 1995, Fowler 1995, Cho 2005, Cole *et al.* 2007) maintain that speakers use more exaggerated articulation, with the effect of greater acoustic/perceptual distinctness, and therefore contrastive focus is realized by a complex of different acoustic parameters that include but are not limited to pitch. We took the approach of including as many measures as possible, although as discussed below, the best performing classifiers turned out to be those using only a handful of these measures.

Phonologists have demonstrated that the mapping between prosodic prominence and contrastive focus is not direct, but involves intermediate, abstract categories like stress and pitch accent. According to “pitch-first” theories (e.g. Bolinger 1958, Pierrehumbert 1980, Selkirk 1995), contrastive focus is realized primarily or uniquely by pitch accents. Recent experimental work (e.g. Rooth 1996, Beaver *et al.* 2007, Howell 2010) has demonstrated that focus may,

under certain pragmatic conditions such as repetition, be realized without pitch, suggesting the possibility of “stress-first” models of contrastive focus.

The rest of this section introduces the machine learning techniques used, their advantages and shortcomings, the manner in which acoustic parameters were selected for inclusion in the machine learning models, the criteria by which we evaluate their success or failure, some of the major finds and their import for linguistic theory.

3.2 Motivation

Two different web harvested corpora of the target *than I did* are reported in this section. The first corpus (*web1*) was collected using an earlier iteration of the harvest methodology, described in Howell & Rooth (2010). The tokens were collected via the *Everyzing* search interface which aggregated podcasts from several content providers. This service went offline in June 2009. The second corpus (*web2*) was collected using a similar methodology, modified for the (now defunct) search interface *multimedia.play.it* with content from CBS Radio and powered with the same technology found in the earlier *Everyzing* interface.

Corpus *web1* contained 91 true tokens of the target *than I did*: 46 tokens with subject focus (category “s”) and 45 tokens with non-subject focus (category “ns”). Corpus *web2* contained 127 true tokens: 62 tokens with subject focus and 65 tokens with non-subject focus.

The *than*-clause and main clause for each token was manually transcribed into English prose. From this transcription, the tokens were manually categorized into one of the two focus categories on the sole criterion, described above, of whether or not both main and *than*-clauses contained the same subject (i.e. whether the subject of the main clause was *I*).

The extraction of acoustic information required annotation at the phonetic level. For each utterance of “than I did”, the following phonetic segments were annotated (cf. Figure 10): V1, the vowel [æ] of *than*; N1, the nasal [n] of *than*; V2, the diphthong [aɪ] of *I*; C3, the stop closure and burst of the initial [d] in *did*; and V3, the vowel [ɪ] of *did*.

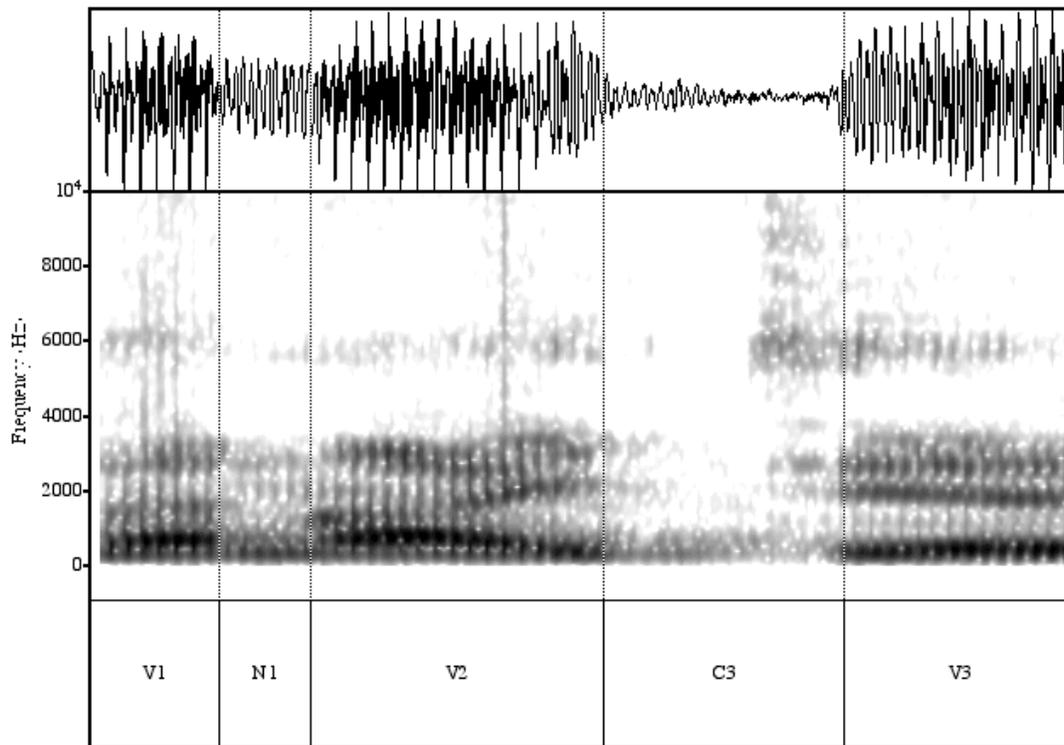


Figure 10. Spectrogram and manual segmentation for one token of ‘than I did’.

A total of 308 acoustic measures were extracted using the scripting function of Praat (Boersma & Weenink 2010). These measures are listed with descriptions in Appendix A. Phenomena of interest included duration, fundamental frequency (f_0), first and second formants (f_1 & f_2), intensity, amplitude, voice quality and spectral tilt. Means or extrema were taken for these phenomena at various loci, such as regular intervals within a vowel or at the time of other extrema.

3.3 Machine Learning

Two machine learning techniques were used to create predictive models of the data. Support vector machines (SVMs) (Boser, Guyon & Vapnik 1992; Cortes & Vapnik 1995) are a relatively recent method of supervised classification that have achieved excellent accuracy in tasks such as object recognition (Evgeniou et al. 2000), cancer morphology identification (Mukherjee et al. 1999) and text categorization (Joachims 1997). Linear discriminant analysis (LDA) (sometimes

known as Fisher linear discriminant analysis after Fisher 1936) has been used widely for several years in pattern recognition tasks.

For both classifiers, a decision divides the space of attributes into two half spaces according to their labels, in our case “subject focus (s)” or “non-subject focus (ns)”. In a dataset with only two sets of attributes the decision function may be represented geometrically as a line dividing a 2-dimensional space (Figure 11), or in a dataset with three sets of attributes, a plane dividing a 3-dimensional space.

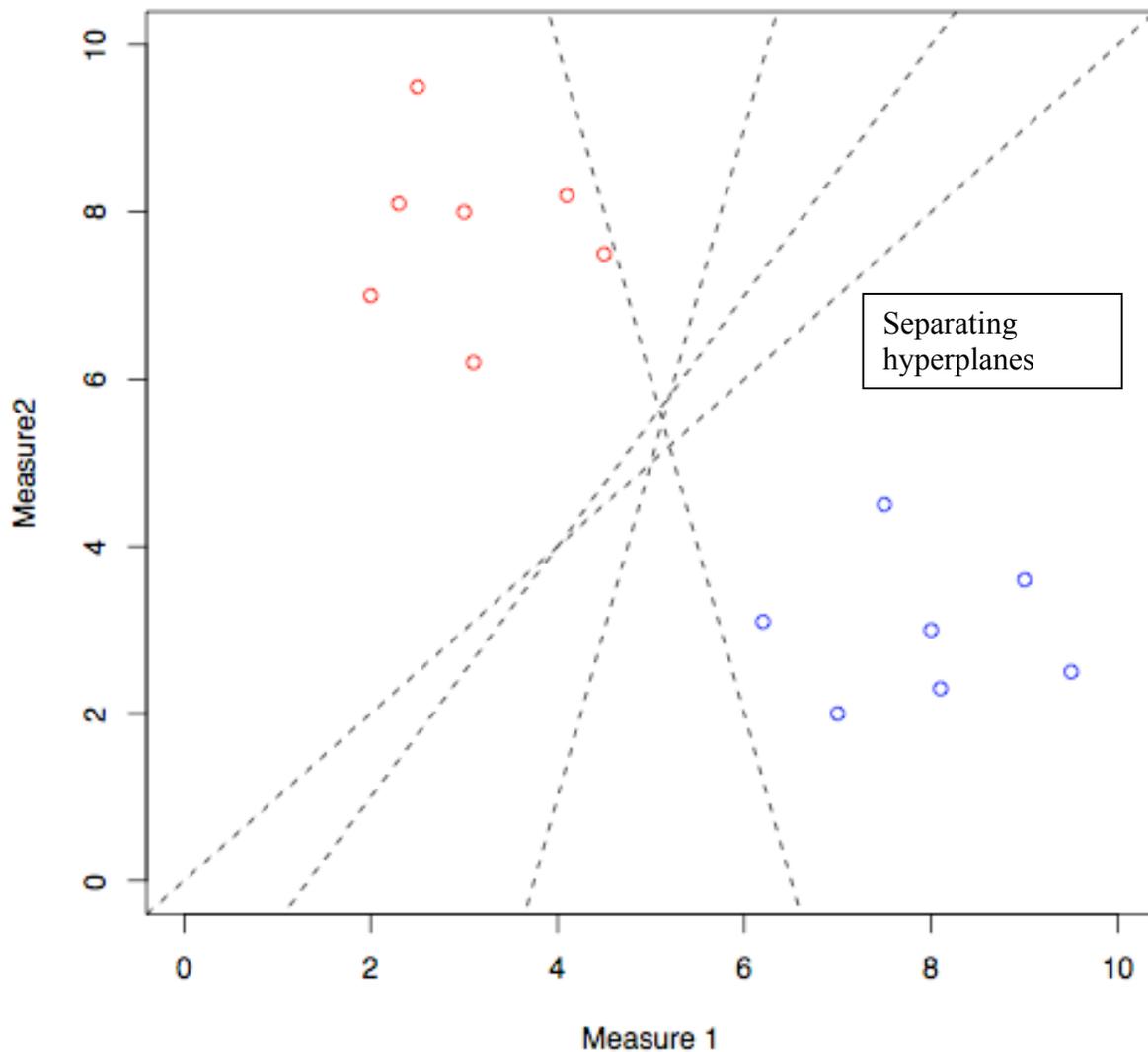


Figure 11. Two dimensional hyperplanes separating binary data.

An SVM classifier looks for the optimal model which maximizes the margin between classes. This approach is considered local, since the optimization is based on data at the boundaries between classes (i.e. the “support vectors”). This is illustrated geometrically in Figure 12 for a two-dimensional space. For this reason, SVM may outperform many conventional classifiers when the number of training data is low and the number of attributes is high. As a maximum margin classifier, SVM also does not assume that the classes are normally distributed or that the classes have equal covariances, although it shares with most classifiers the assumption that the training and test data are independent and produced in the same way.

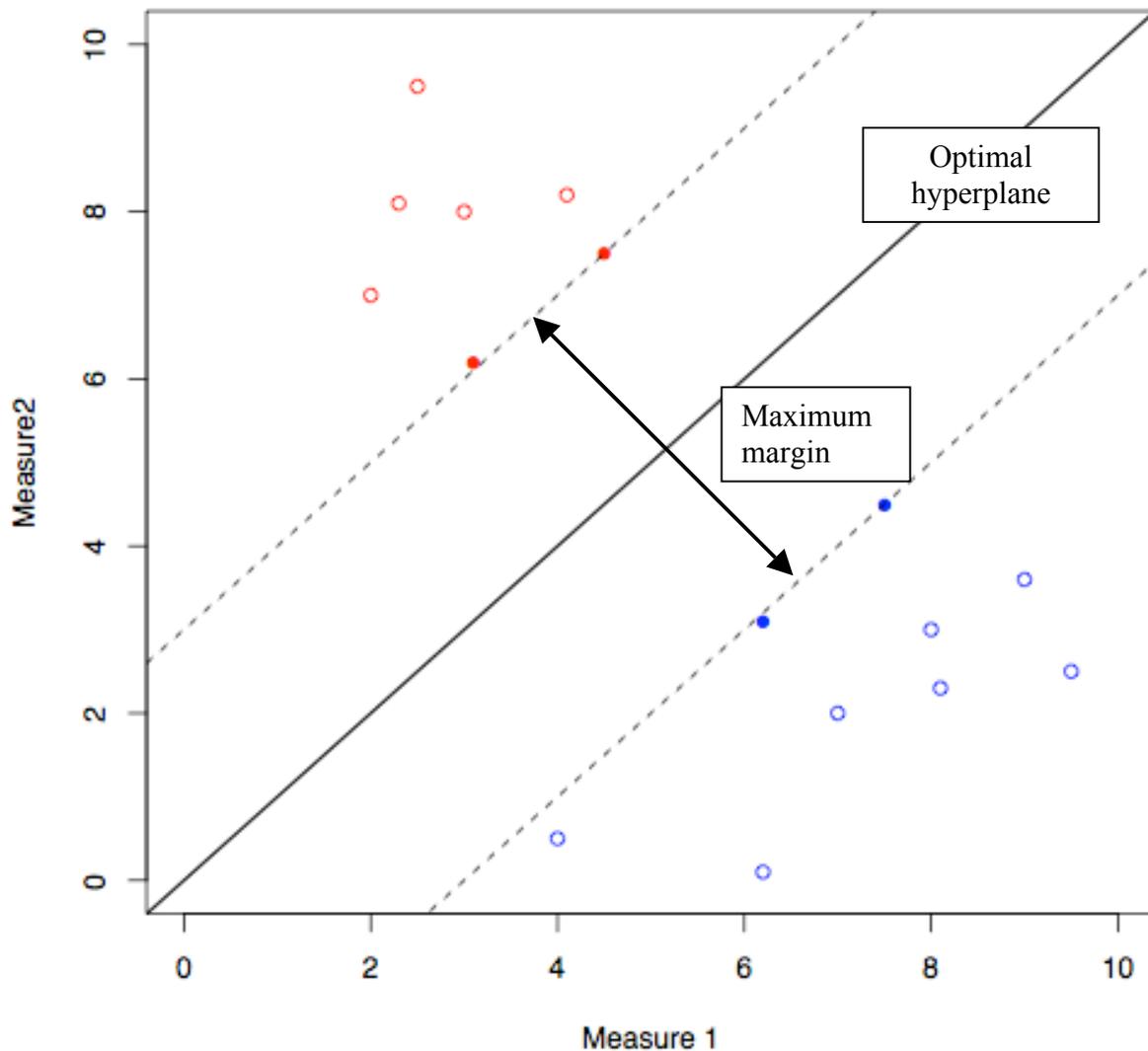


Figure 12. Support Vector Machine (SVM). Optimal two-dimensional hyperplane and maximum margins separating binary data.

Another feature of SVMs is the mapping of linear attributes into a multi-dimensional feature space, the so-called “kernel trick”. By expressing the decision function in dual coordinates, it is possible to introduce a kernel function. This greatly reduces the complexity of the algorithm and allows it to scale well with a large number of examples. Although the data should be internally

scaled for best results, use of a non-linear kernel also avoids the need to transform attributes which may be non-linear, such as duration or acoustic energy.

Many kernel functions have been used successfully in different classification tasks. Hsu *et al.* (2003) recommend a radial basis function (RBF), a non-linear mapping which has been shown to also encompass a linear kernel (Keerthi 2003) and behave similarly to a sigmoid kernel (Lin & Lin 2003). Hsu *et al.* note that the RBF kernel requires only two hyperparameters, while a polynomial kernel, for example, will contain two or more, contributing to model complexity. (All kernels contain at least one hyperparameter C , cost or constant.) At the same time, Hsu *et al.* also suggest that the results of a linear kernel may be comparable with those of an RBF kernel in situations where the number of attributes to be mapped is greater than the number of data instances, a situation which obtains with a full model of the *web1* dataset. We consider both RBF and linear kernels. The implementation of SVM used here comes from the *libsvm* package (Chang & Lin 2001) for R.

An LDA classifier looks for the optimal model which minimizes within-class distance and maximizes between-class distance. This approach is considered global, since the optimization is based on the mean and covariance of the classes, which are usually obtained via a discriminant function of ordinary least-squares or maximum likelihood estimation¹. LDA makes many assumptions, including normal distribution of classes and homogeneity of covariances. Classes in the *web1* and *web2* datasets of this chapter are well balanced, although it is unlikely that the variances of all 308 attributes are normally distributed. Poor results may also obtain if the training set is small. Furthermore, the LDA classifier has been shown to perform best when the number of attributes is minimized (ideally no greater than 2 attributes for a binary classifier) and the attributes are not intercorrelated (cf. Brown & Wicker 2000). In practice, however, it is often possible to obtain good results even with small datasets and data which violate the assumptions of normal distribution and homogeneity of covariances (e.g. Lachenbruch 1975; Klecka 1980; Stevens 2002). The implementation of linear discriminant function analysis used here comes from the MASS package (Venables & Ripley 2002) for the statistical computing environment R (R Development Core Team 2008).

¹ The R function `lda` in R package MASS is described by its authors in Venables & Ripley 2002 *Modern Applied Statistics with S*, 4th ed., pp. 331-334.

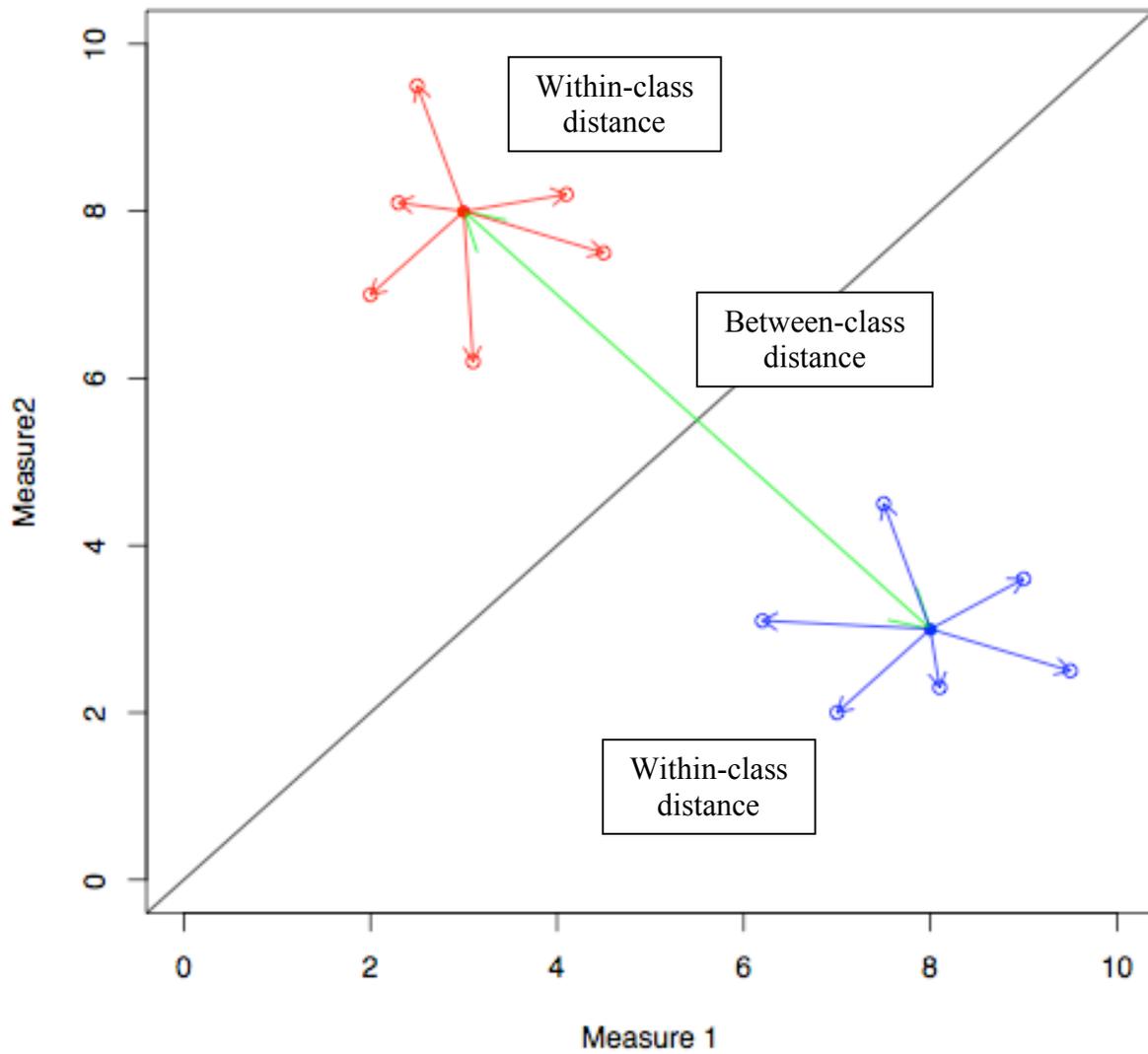


Figure 13. Linear Discriminant Analysis (LDA). Optimal two-dimensional hyperplane and between-class and within-class distances for binary data.

3.4 Feature Selection

Feature selection is a necessity for LDA where the possibility of collinear features exists. Indeed, the R implementation of LDA is halted and cannot proceed in case of high collinearity. As for SVM, one reason to use this classifier is precisely to avoid costly feature selection; nonetheless, feature selection prior to or in the process of building an SVM classifier has been shown to improve the generalization accuracy and/or model complexity (and thus computation) for those datasets with redundant or irrelevant features.

Feature selection is also a means of peering into the “black box”, and understanding which features are contributing to a model’s generalization accuracy. For example, a classifier which accurately predicts a focus category may be the goal, but we also wish to know which acoustic measures are important for this task. The set of acoustic measures used by a classifier to predict focus are not necessarily equivalent to the set of acoustic measures that an individual human listener may use in the same classification task, but the question of whether and why the machine-learning and human sets of attributes are not equivalent is in fact a useful research question provided by the classifier.

Most authors agree that some combination of manual and statistical feature selection techniques may be used, although there is no consensus on the ordering or relative importance of manual or statistical feature selection. We used both automatic feature selection and manual feature selection, in some cases informed by theoretical expectations and in some cases through basic trial-and-error.

We used a feature selection algorithm VarSelRF (Diaz-Uriarte 2009), which is designed for genetic research, in which datasets typically contain large sets of features for relatively few data instances. This algorithm based on a random forests method of classification and uses backwards variable elimination. This is a “filter” method of feature selection, since it occurs as a kind of preprocessing before a model is trained. When applied to all 308 features, this algorithm selected the following four features in *Automated Feature Selection A*.

(2) Automated Feature Selection A (from all 308 features)

<i>dur_V2</i>	duration of <i>I</i>
<i>mean_f0_ratio</i>	ratio of mean fundamental frequency (cf. pitch) in <i>I</i> and <i>did</i>
<i>f1f2_40_V2</i>	difference between first and second formant values (i.e. vocal tract resonances) in <i>I</i> , measured at 40% into the vowel

flf2_50_V2 difference between first and second formant values in *I*, measured at 50% into the vowel

Classifiers using the experimenter-selected feature sets turned out to perform better than those using the automated feature sets, although the automation process helped to inform the manual selection, since trial-and-error with 308 measures was of course not feasible. One of the best-performing experimenter-selected feature sets contained a different but overlapping set of four features.

(3) Experimenter-Selected Feature Selection A

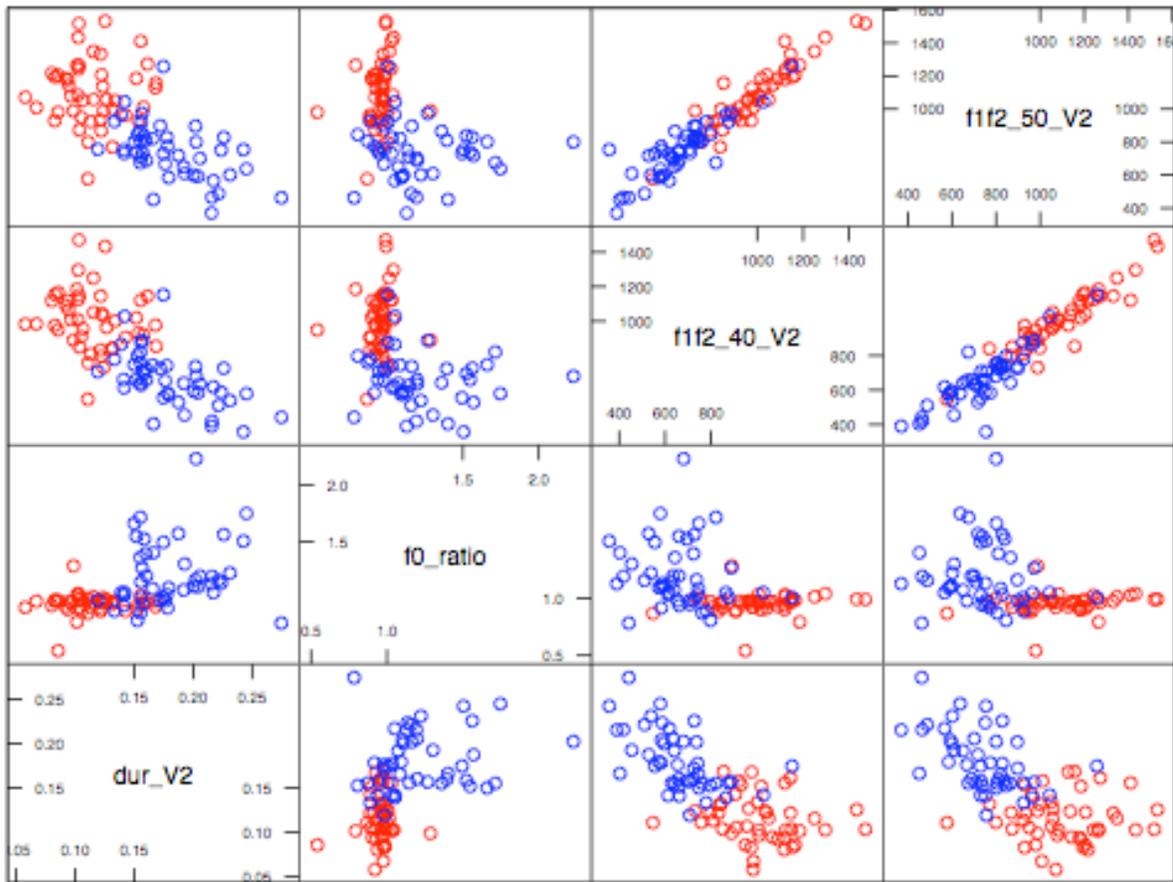
dur_V2 duration of *I*
dur_C3 duration of first stop closure (i.e. the silence corresponding to the tongue constriction) in *did*
mean_f0_ratio ratio of mean fundamental frequency (cf. pitch) in *I* and *did*
flf2_50_V2 difference between first and second formant values in *I*, measured at 50% into the vowel

Several other feature combinations were evaluated (discussed in Howell 2011); only those in (2) and (3) are considered here, for ease of presentation.

Representing a four-dimensional classifier graphically is difficult, but it is possible to get a good visual sense of the separation provided by these measures with scatter plot matrices, as in Figures 14 and 15.

Automated Feature Selection A

○ s ○ ns



Scatter Plot Matrix

Figure 14. Pairwise comparison of the *web1* data for the four features from Automated Feature Selection A.

Experimenter-Selected Feature Selection A

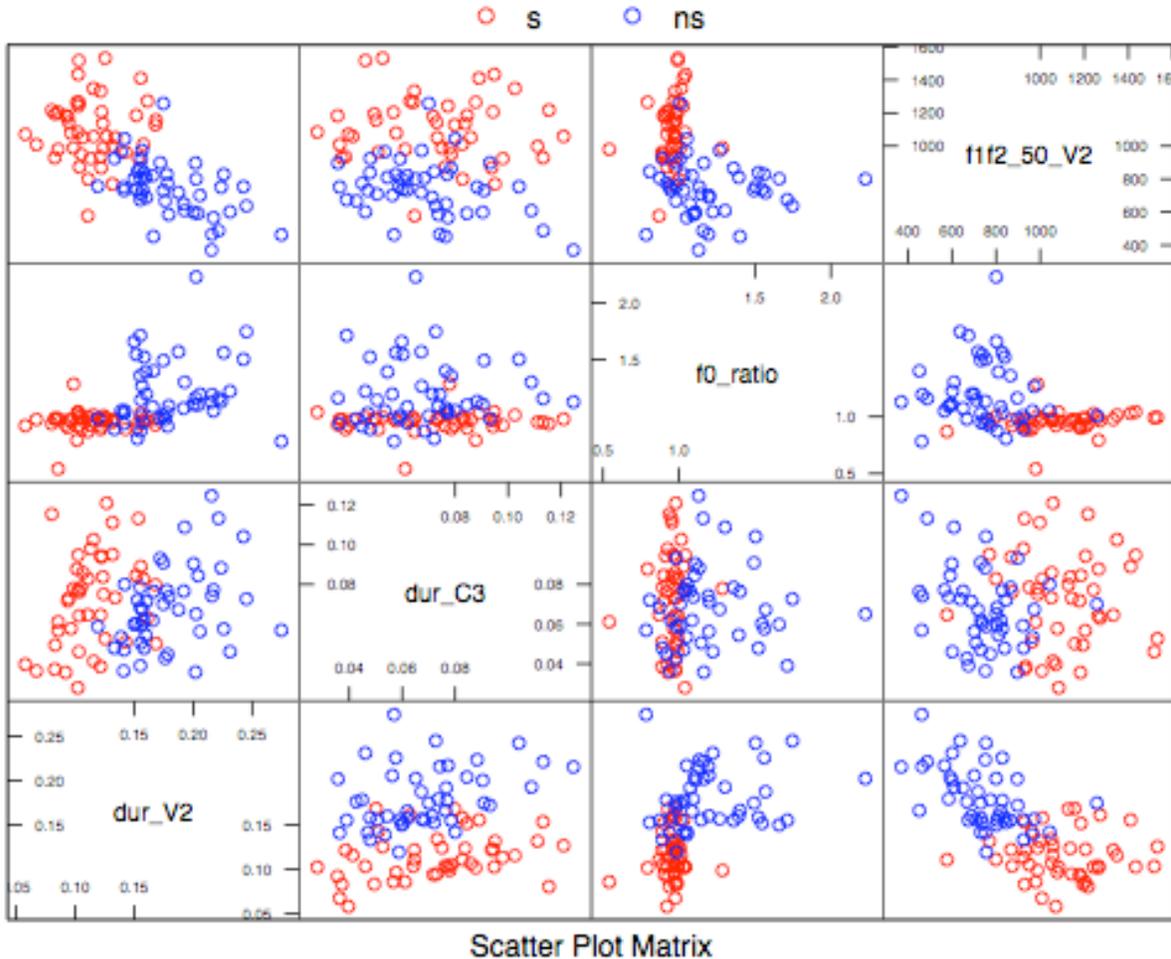


Figure 15. Pairwise comparison of the *web1* data for the four features from Experimenter-Selected Feature Selection A.

3.5 Evaluation

Following convention in the machine learning community, we evaluated our classifiers by training them on one set of data (*web1*) and testing them on a new set of data (*web2*). The percentage of correctly classified tokens is termed the generalization accuracy rate. This is compared against a baseline accuracy, which is simply the percentage of the most frequently

occurring category (“s” or “ns”). Finally, a balanced error rate takes into consideration the relative number of false positives and false negatives.

(3) Baseline accuracy

$$\frac{\text{\# tokens in largest class of test set}}{\text{\# tokens in both classes in test set}}$$

(4) Generalization accuracy

$$\frac{\text{\# of tokens in test set accurately classified}}{\text{\# of tokens in test set}}$$

(5) Balanced error rate

$$\frac{\text{\# incorrect "s"}}{\text{\# total "s"}} * \frac{\text{\# incorrect "ns"}}{\text{\# total "ns"}} * \frac{1}{2} * 100$$

3.6 Results

All of the models trained on web corpus dataset *web1* achieved generalization accuracy and balanced error rate on the second web corpus dataset *web2* well above the baseline (accuracy 51.2/ BER 48.8). The different machine learning methods and feature sets were also quite competitive with each other. A summary of results is listed in Figure 16.

	Classifier performance on <i>web2</i>			
Feature set	Baseline	SVM (radial kernel)	SVM (linear kernel)	LDA
1. Full set (no. features = 308)	51.2	82.7 (14.3)	85.0 (14.2)	n/a due to collinearity
2. Automated feature selection A (no. features = 16)	51.2	89.8 (10.1)	89.0 (10.3)	90.6 (9.0)
3. Automated feature selection B (no. features = 4)	51.2	86.6 (13.1)	90.6 (9.2)	87.4 (11.9)
4. Experimenter-selected A (no. features = 4)	51.2	92.9 (6.5)	92.1 (7.1)	91.3 (7.7)
5. Experimenter-selected B (no. features = 3)	51.2	91.3 (7.7)	92.1 (7.1)	91.3 (7.7)

Figure 16. Generalization accuracy rates (and balanced error rates) for different machine learning models trained on *web1* and tested on *web2*.

3.7 Discussion

Three observations from these results are of particular note. First, the results overwhelmingly confirm the theoretical predictions for the *location* of acoustic prominence. The machine learning classifiers can predict the semantic categories from acoustic information alone with considerable accuracy. As for those tokens which were erroneously classified, four were significant predictors of the human classifiers' performance, with mean accuracies of 59.0, 28.2, 23.1 and 17.9 percent. On further inspection, prominence in these tokens, predicted to be greatest on the subject because of a distinct, contrasting subject in the main clause, may plausibly be explained by other factors such as speech disfluency or extralinguistic emphasis.

Second, classifiers with information about fundamental frequency (*f0*) performed on par or better than classifiers which lacked information about fundamental frequency. Experimenter-

selected feature set A contained the features *dur_V2*, *dur_C3*, *f1f2_50_V2* and *mean_f0_ratio*; experimenter-selected feature set B contained the first three, but lacked *mean_f0_ratio*. Third, all of the best-performing classifiers used measures of duration and measures of first and second formant differences in the vowel of *I*.

The predictiveness of the durational and vowel quality features, and the apparent non-predictiveness of *f0* features, are consistent with hyperarticulation theories of contrastive focus realization. According to this theory, the articulation of *I* is exaggerated when contrastively focused, a speaker taking more time to pronounce it; this extra time also allows formant targets for the vowels to fully realized or even overshoot. It also follows that a fundamental frequency target may also be fully realized, resulting in a higher peak, although on this account the increased *f0* peak would be just one of several acoustic cues, rather than the primary cue.

On the other hand, pitch-first theories generally require pitch accents overlaid on sentence-level stress, and sentence-level stress is realized by acoustic cues of duration and vowel quality. So it is plausible that, in the absence of accurate *f0* measurements—algorithms for extracting *f0* are notoriously fallible—acoustic cues of sentence-level stress are the next-best set of predictors.

The scientific literature on acoustic prominence is dominated by discussion of fundamental frequency. (Kochanski 2006 reports that articles about *f0* outnumber articles investigating other cues by nearly 5 to 1.) The predictiveness of duration and vowel quality and the apparent non-predictiveness of *f0* is therefore quite significant.

Finally, it is traditionally held that prominence is “syntagmatic”, meaning that prominence is processed relative to the sentence that is being uttered (e.g. Jakobson, Fant and Halle 1951; Trubetzkoy 1935,1939; Lehiste 1970; Ladefoged 1975; Hyman 1978). This explains, for example, how a word may be perceived as prominent in either fast or slow speech. Ladd (1991,1996) argues, however, that prominence in an utterance is also processed “paradigmatically”, relative to ways the utterance could have been produced. Many of the measures we used were syntagmatic, such as ratios of measures taken from *I* and the vowel in *did*. Strikingly, the best-performing classifiers did not use these syntagmatic measures, but instead used measures taken from one or other of the vowels. This suggests that listeners may be using more paradigmatic information than previously assumed. From a practical standpoint, it

also suggests a more efficient automatic detection of focus which is localized to the word or syllable level.

4. Lab Experiments

4.1 Cross-validation of Harvest and Laboratory Data

Most data linguistic theories are based on are based on introspective intuition or on data collected in the lab. Lab experiments are usually scripted and are usually very close to reading, and far from spontaneous speech. Experiments often reveal the interest of the researcher since they contain an untypical proportion of certain examples of interest. In other words, the generalizability of much of our knowledge of the acoustic correlates of linguistically relevant factors is questionable. Cross-validation of laboratory results with more naturalistic and spontaneously produced data could solve this methodological problem, but such data is hard to come by.

There is of course a field of corpus linguistics which works with data that were produced under more natural circumstances, for example the Switchboard Corpus (Godfrey *et al.* 1992), which is available through the Linguistics Data Consortium. It consists of conversations between two people whose task was to schedule a meeting over the phone. This type of data is more naturalistic than typical experimental data since it involves a rather unconstrained conversation between two people. However, the corpus is not particularly big, and many types of constructions of linguistic interest will not appear at all. The same issue of scale applies to other spoken corpora.

Our harvest procedure can fill this gap in our methodological toolbox: The amount of data available online is vast, and if we find a way to systematically harvest data sets of interest, then we can complement lab-results with real-life data, and thus cross-validate data collected in a controlled environment with comparable data collected in a much noisier and much more variable channel.

This cross-validation cuts both ways: results obtained under the more controlled conditions in a lab-environment can inform our analysis of naturally-occurring data, where there is far bigger variability due to differences in recording conditions, differences in register and levels of formality. The lab data can be used as a standard of comparison and a guide in developing

acoustic measures to be extracted from corpus data. Conversely, the new data from harvested data sets will provide a way to validate results obtained in lab environments. Sources of variation that are either excluded in lab experiments or controlled for (speaker and item effects), may lead to reliable results that turn out to be irrelevant when looking at real world data, due to the much broader range of factors influencing the signal.

4.2 Matlab Scripts to Conduct Production and Perception Experiments

A suite of matlab scripts was developed, that provides a platform to conduct production and perception experiments. In production experiments, we collect particular data points of interest by prompting participants to say them out loud as naturally as possible. In perception experiments we ask people to rate acceptability, or to annotate the data for prominence based on their own intuition (Cole), or we conduct a ‘context-retrieval’ task (Gussenhoven 1983).

The production experiments are scripted and the resulting data can sound monotonous since participants fall into a reading pattern can sound quite unlike spontaneous speech. There are several options that help avoid such a drab reading intonation: The participant can be asked to memorize the sentence, and on the recording screen the sentence is no longer visible. This way, the sentence has to be produced from memory, which makes it less likely that the participant will read the sentence off the screen without really processing it. It is also possible to conduct pseudo-dialogues: We pre-record the part of the interlocutor. Participants see the entire dialogue, and then press a key when they’re ready. They then hear the part the other person says played through their headset, and have to respond their part as naturally as possible. In our experience, this works well in prompting more natural productions.

The input to the matlab scripts is a simple tab-delimited spread sheet, with the data organized into items and conditions. The script then creates a pseudo-randomized playlist drawn from the list. There are several possible designs: Latin square, where every participant sees one condition from each item, but an equal number of trials from each condition over the entire experiment. Between subjects, were every participant only sees the same condition from each item. Or Within subjects, were every participant sees every condition from each item in a pseudo-randomized order, such that no condition is repeated more than once and there are no adjacent trials from the same item.

The recorded data is then ‘filtered’, which means that an RA goes through them and checks whether the participant said exactly the utterance that she was scripted to say. If there is a slight deviation either the text transcript can be changed by the RA, or the utterance can be marked as ‘problematic’ and excluded from analysis. The filtering is done by a praat script similar to the one we use for filtering the harvest data. The filtering is the only manual step in the analysis process.

After filtering, the data is forced aligned using the HTK force aligner (we are using a set of scripts made available by Kyle Gorman (UPenn), which have since been superseded by the set of Python scripts underlying the Penn-Forced-Aligner. So far, we have trained our forced aligner on 10 hours of lab speech collected in our own lab, while the Penn-Forced Aligner is trained on corpus data. The scripts look up a phonetic transcription of the utterance in the CMU pronunciation dictionary, and then goes through various rounds of estimating the best alignment between transcription and sound file. The output of the forced-aligner are praat-textgrids which contain annotation tiers for a segment-by-segment and a word-by-word annotation.

After forced-alignment, additional annotations are added to the textgrids. The tab-delimited experiment file consists a column which indicates the ‘words of interest’ of every utterance. This column contains the text of the utterance in which words in whose acoustic properties we’re interested in are marked. This information is used to introduce a special ‘word of interest’ tier in the textgrids. We can then use a Praat script to extract various acoustic measures for each word of interest and subject this to acoustic analysis.

This experimental pipe-line is working quite effectively now, and we have already conducted many experiments using it. The remainder of this section summarizes our first complete series of experiments cross-validating a harvest result.

4.3 Than-I-Did Experiments

In the past 6 months, after our matlab scripts and the associated HTK forced-aligner were ready to go, we have conducted a series of cross-validation experiments on our first data set, the ‘than I did’ set. In total, we conducted 3 experiments.

4.3.1 Laboratory Experiment 1: Naive Prominence Annotation of Harvest Data (Perception)

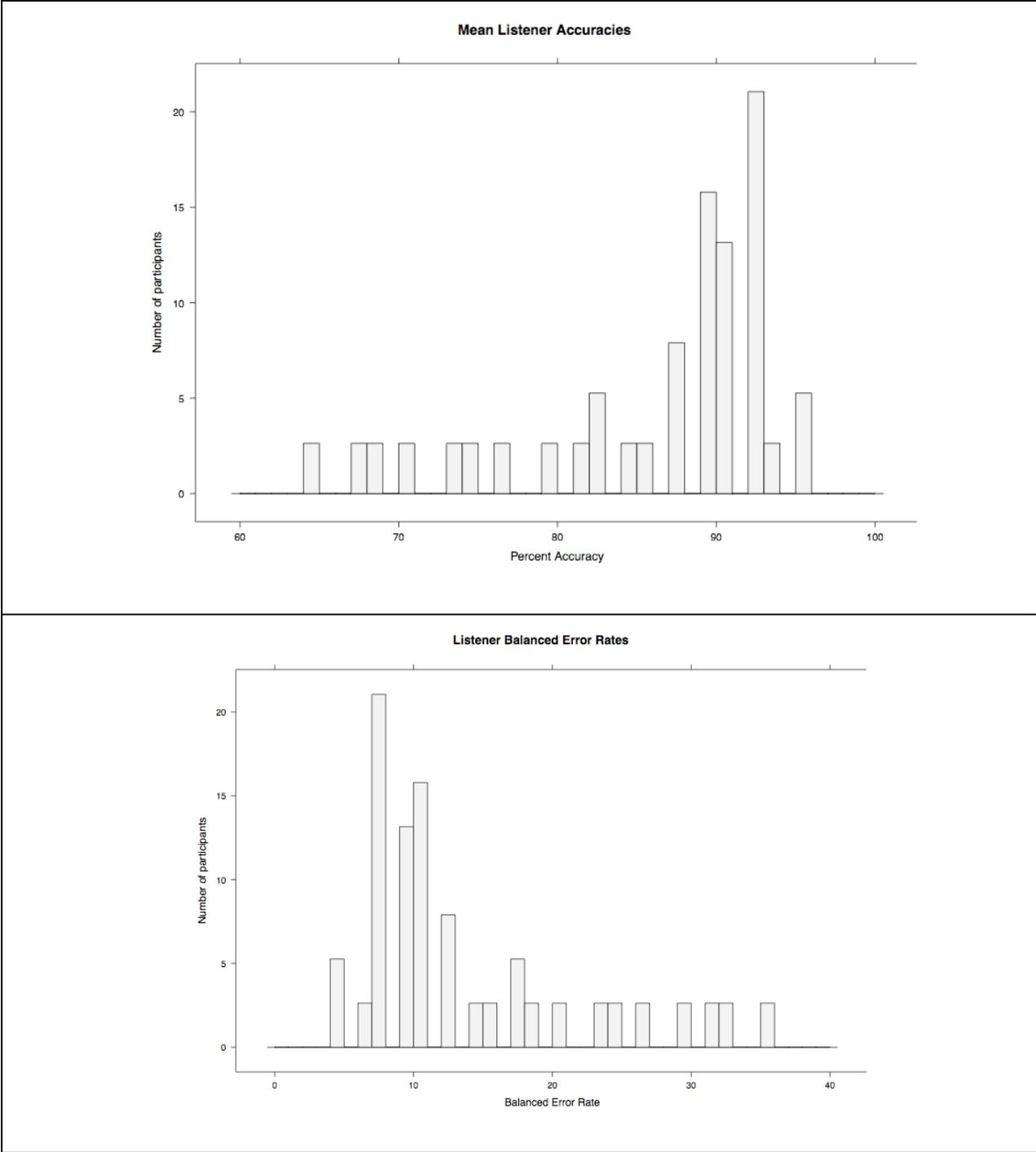
In order to evaluate the performance of the machine learning classifiers, we conducted a “human classifier” experiment using a subset of the harvested speech data. A subset of 64 tokens from the *web2* corpus dataset was chosen: the first 32 of each focus category. From each soundfile, the sequence “than I did” was extracted. The information presented to participants of the perception experiment was limited in this way in order to more closely replicate the limited information available to the machine learning algorithms: neither the machine classifier nor the human listener had the preceding or following acoustic information, nor did they have any grammatical or pragmatic context.

Forty individuals participated in the perception experiment, which was conducted at the prosody lab at McGill University. Participants were compensated for their time. The data of two participants was not analyzed because the subjects reported making errors. The stimuli were played one at a time, in random order, with no category repeated more than twice. After each stimuli, the listener was asked to complete two tasks: first, to choose whether “I” or “did” had greater prominence; second, to rate confidence in their choice on a scale from 1 (“very confident”) to 5 (“very uncertain”). Participants’ confidence rating turned out to be a very significant predictor of their performance on a given stimuli (generalized linear model: $\sigma = 0.031$, $z = -10.81$, $p < 0.001$).

(6) Questions elicited in laboratory perception experiment

- Question 1: Which is more prominent: I or did?
Question 2: How confident are you?
(very uncertain) 1 2 3 4 5 6 7 (very confident)

The human acoustic classifiers performed on par with the machine learning classifiers. The 38 listeners in the perceptual experiment achieved a mean accuracy of 85.9 percent, median accuracy of 89.1 percent and balanced error rate of 14.1 percent. Participants’ individual accuracy rates ranged from 64.1 to 95.3 percent and their balanced error rates ranged from 4.7 to 35.9 percent.



Figures 17 and 18. Distributions of listener accuracies and balanced error rates in Laboratory Experiment 1.

The features which were most predictive in the machine learning experiments were also significant predictors of listeners' responses. In generalized linear mixed models that

incorporated speaker and item as random effects (see Figure 19), there were main effects for each of the acoustic variables, with the notable exception of *mean f0 ratio* in the model using experiment selection A. There were no effects for participant or item.

EXPERIMENTER SELECTION A: duration_V2, duration_C3, flf2Time50_V2, f0_ratio

Random effects:

Groups	Variance	Std. Dev.
Participant	0.041699	0.20420
Item	0.033360	0.18265

Fixed effects

	Estimate	Std. Error	z-value	p-value
<i>Intercept</i>	1.289440	0.514789	2.505	0.0123 *
duration_V2	36.567667	2.049602	17.841	<2e-16 *
duration_C3	-45.726612	3.095865	-14.770	<2e-16 *
flf2Time50_V2	-0.003150	0.000293	-10.749	<2e-16 *
mean_f0_ratio	-0.062636	0.235012	-0.267	0.7898 n.s.

EXPERIMENTER-SELECTED B: duration_V2, duration_C3, flf2Time50_V2

Random effects:

Groups	Variance	Std. Dev.
Participant	0.041699	0.20420
Item	0.033360	0.18265

Fixed effects

	Estimate	Std. Error	z-value	p-value
<i>Intercept</i>	1.231643	0.463617	2.657	0.0079 *
duration_V2	36.415015	1.977036	18.419	<2e-16 *
duration_C3	-45.492161	2.971366	-15.310	<2e-16 *
flf2Time50_V2	-0.003152	0.000293	-10.758	<2e-16 *

Figure 19. Summary of generalized linear mixed models for listener responses to a subset of the web corpus data using predictors from the hand-selected feature sets. Statistical significance ($p < 0.01$) indicated by asterisks.

4.3.2 Experiment 2: Cross-Validation (Production)

In order to compare natural speech found in the web data to speech elicited in the laboratory, we conducted a speech production experiment. 16 written stimuli containing the target *than I did* were constructed by the experimenters based on actual tokens from the web corpus. The 16 stimuli were further divided into different experimental conditions, such as question and statement (discussed in Howell 2011 but omitted here for space), and were balanced between the two focus categories, “s” and “ns”. 27 individuals participated, although one participant’s speech failed to be recorded, leaving a total of 26 participants.

The written stimuli were presented to participants on a computer screen. After reading the text aloud, participants were asked to rate the naturalness of the written stimuli on a scale from 1 (very natural) to 5 (very awkward). The mean rating for the individual stimuli ranged from 1.72 to 3.08; the overall mean was 2.35. 19 recordings were discarded due to disfluencies, such as false starts, hesitations or utterances that did not match the written stimuli. Automatic alignment failed on 3 files, leaving a total of 394 usable tokens.

We applied the same machine learning classifiers which were trained on the harvested data (*web1*) to the laboratory-elicited data. As before, classifiers using the experimenter-selected feature sets turned out to perform better than those using the automated feature sets. The generalized accuracy rates and balanced error rates were somewhat lower than those achieved by the web-trained/web-tested classifiers, but considerably higher than the baseline and on par with the human performance rates.

Web-trained, lab-tested classifiers

Feature set	Baseline	lab		
		SVM (RBF)	SVM (linear)	LDA
1. Full (no. features = 308)	51.0	72.3 (19.9)	82.2 (16.3)	--
2. Automated feature selection A (no. features = 16)	51.0	81.5 (18.1)	84.5 (14.4)	74.4 (21.8)
3. Automated features selection B (no. features = 4)	51.0	83.8 (16.1)	74.1 (20.9)	71.3 (21.9)
4. Experimenter-selected A (no. features = 4)	51.0	82.2 (16.8)	74.6 (21.2)	75.1 (19.4)
5. Experimenter-selected B (no. features = 3)	51.0	85.8 (11.9)	87.3 (11.0)	84.3 (12.9)

Figure 20. Generalization accuracy rates (and balanced error rates) for different machine learning models trained on *web1* & *web2*, and tested on *lab*.

Finally, we trained machine learning classifiers on the laboratory data (*lab*) and tested them on the web data (*web1* and *web 2* collectively). We used the same algorithm *VarSelRF* for the automated feature selection. Note that because the *VarSelRF* algorithm was applied to a different training set (i.e. the laboratory data), instead of the harvested data, it yielded slightly different feature sets. The experimenter-selected feature sets are the same. (Again, see Howell 2011 for a more exhaustive list of the models considered.)

Lab-trained, web-tested classifiers

		<u><i>web1</i> & <i>web2</i></u>		
Feature set	Baseline	SVM (RBF)	SVM (linear)	LDA
1. Full set (no. features = 308)	50.5	79.8 (17.4)	73.4 (24.7)	--
2. Automated feature selection C (no. features = 43)	50.5	83.9 (15.2)	79.4 (20.6)	--
3. Automated feature selection D (no. features = 18)	50.5	81.7 (16.8)	72.9 (27.1)	--
4. Experimenter-selected A (no. features = 3)	50.5	89.4 (9.8)	88.5 (10.3)	88.1 (10.9)
5. Experimenter-selected B (no. features = 4)	50.5	85.8 (12.9)	88.5 (10.3)	88.1 (10.9)

Figure 21. Generalization accuracy rates (and balanced error rates) for different machine learning models trained on *lab*, and tested on *web1* & *web2*.

Again, classifiers using the experimenter-selected feature sets turned out to perform better than those using the automated feature sets. The generalized accuracy rates and balanced error rates were lower than those achieved by the web-trained and web-tested classifiers, but considerably higher than the baseline and on par with the human performance rates.

The results of these machine learning experiments confirm the data collected in the laboratory are sufficiently representative of naturally-occurring speech. The theoretical

predictions for focus *placement* hold for both corpus and laboratory datasets and both datasets support theories of focus *realization* in which pitch accents and fundamental frequency are not the sole correlates of focus.

4.3.3 Experiment 3: Naive Prominence Annotation of Lab Data (Perception)

In the third experiment, human listeners were presented with excerpts of “than I did” taken from the laboratory production data recorded in experiment 2. The experiment was carried out with the same methodology used in Experiment 1. Forty-one individuals participated. Participants’ confidence rating turned out to be a very significant predictor of their performance on a given stimuli (generalized mixed-effects linear model: $\sigma = 0.05844$, $z = 7.429$, $p < 0.001$).

The human acoustic classifiers performed on par with the machine learning classifiers. The 41 listeners in the perceptual experiment achieved a mean accuracy of 78.5 percent, median accuracy of 81.3 percent and mean balanced error rate of 13.1 percent. Participants’ individual accuracy rates ranged from 53.1 to 96.9 percent and their balanced error rates ranged from 3.7 to 29.3 percent.

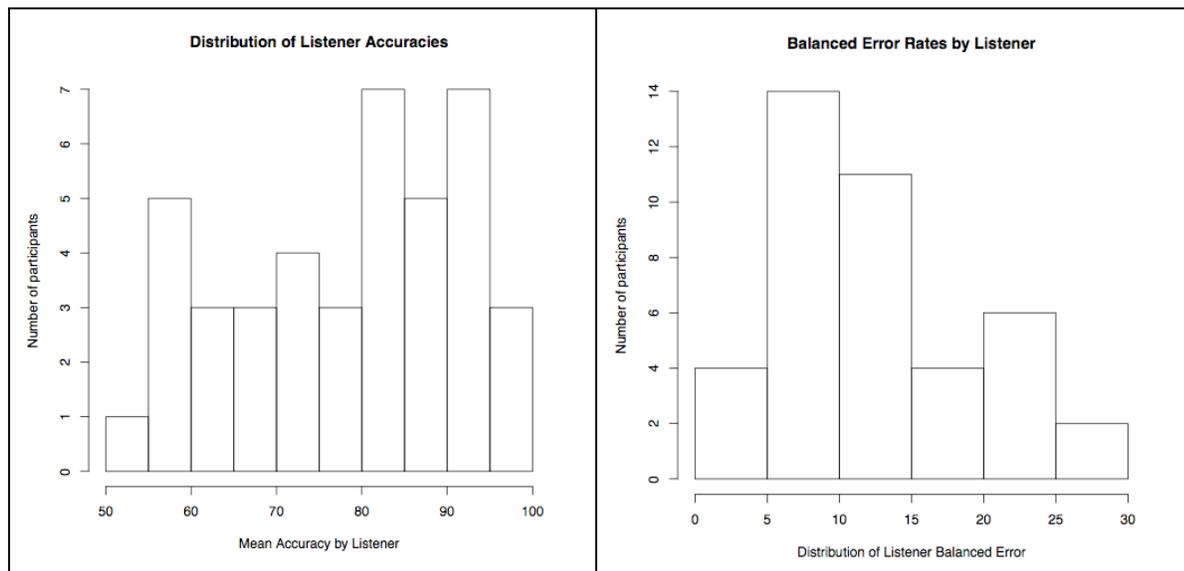


Figure 22. Distributions of listener accuracies and balanced error rates.

The features which were most predictive in the machine learning experiments were also significant predictors of listeners' responses. In generalized linear mixed models that incorporated speaker, listener and item as random effects (see Figure 23), there were main effects for each of the acoustic variables. There were no effects for participant, speaker or item.

EXPERIMENTER SELECTED A: duration_V2, duration_C3, f1f2Time50_V2, f0_ratio

Random effects:

Groups	Variance	Std. Dev.
Participant	0.0834607	0.288896
Speaker	0.0162850	0.127613
Item	0.0081426	0.090236

Fixed effects

	Estimate	Std. Error	z-value	p-value
<i>Intercept</i>	5.182e+00	1.114e+00	4.650	3.33e-06 *
duration_V2	1.447e+01	3.919e+00	3.692	0.000222 *
duration_C3	-1.821e+01	8.370e+00	-2.176	0.029581 *
f1f2Time50_V2	-5.662e-03	6.301e-04	-8.985	< 2e-16 *
mean_f0_ratio	4.082e-01	1.498e-01	2.724	0.006448 *

EXPERIMENTER-SELECTED B: duration_V2, duration_C3, f1f2Time50_V2

Random effects:

Groups	Variance	Std. Dev.
Participant	0.0834607	0.288896
Speaker	0.0162850	0.127613
Item	0.0081426	0.090236

Fixed effects

	Estimate	Std. Error	z-value	p-value
<i>Intercept</i>	5.261e+00	1.123e+00	4.685	2.80e-06 *
duration_V2	1.398e+01	4.004e+00	3.492	0.00048 *
duration_C3	-1.742e+01	8.470e+00	-2.057	0.03969 *
f1f2Time50_V2	-5.367e-03	6.145e-04	-8.734	< 2e-16 *

Figure 23. Summary of generalized linear mixed models for listener responses to a subset of the laboratory-elicited production data using predictors from the hand-selected feature sets. Statistical significance ($p < 0.01$) indicated by asterisks.

The high performance of the machine learning classifiers demonstrate that they can mimic human behavior. The evidence from the perception experiments that humans use the same sets of acoustic features suggests that the machine learning classifiers are useful representations of human behavior.

References

- Altenberg, Bengt. 1987. Prosodic Patterns in Spoken English: Studies in the Correlation Between Prosody and Grammar for Text-to-Speech Conversion, Lund studies in English; 76, Lund University Press, Lund, Sweden.
- Beaver, David I., Brady Z. Clark, Edward Flemming, T. Florian Jaeger and Maria Wolters. 2007. When Semantics Meets Phonetics: Acoustical Studies of Second-Occurrence Focus. *Language* 8: 245-276.
- Boersma, Paul and David Weenink. 2010. Praat, a system for doing phonetics by computer. *Glott International* 5:341–345.
- Bolinger, Dwight. 1958. Stress and Information. *American Speech* 33:5-20.
- Boser, B. E, I. M Guyon, and V. N Vapnik. 1992. A training algorithm for optimal margin classifiers. P. 144–152 in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*.
- Brenier, Jason. 2008. *The Automatic Prediction of Prosodic Prominence from Text*. Doctoral dissertation. University of Colorado at Boulder.
- Brown, Michael T. and Lori R. Wicker. 2000. Discriminant Analysis. In *Handbook of applied multivariate statistics and mathematical modeling*, edited by Howard Tinsley pp. 209-235. San Diego: Academic Press.
- Campbell, Nick and Mary Beckman. 1997. Stress, prominence, and spectral tilt. In *Intonation: Theory, models and applications (proceedings of an ESCA workshop)*, September 18-20, 1997, Athens, Greece), ed. by George Carayiannis, Antonis Botinis and Georgios Kouroupetroglou. ESCA and University of Athens Department of Informatics.

- Chang, Chih and Chih Lin. 2001. LIBSVM: a library for support vector machines. URL <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Ken Chen and Mark Hasegawa-Johnson. 2004. How prosody improves word recognition. *SpeechProsody 2004*, Nara, Japan.
- Cho, Taehong. 2005. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /α,i/ in English. *Journal of the Acoustical Society of America* 117:3867-3878.
- Cole, Jennifer, Kim, H., Choi, H. and Hasegawa-Johnson, Mark. 2007. Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics* 35: 180-209.
- Conkie, A., Riccardi, G. & Rose, R. C. 1999. Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events. In *Proceedings of Eurospeech*, pp. 523–526. Budapest, Hungary.
- Cortes, Corina and Vladimir Vapnik. 1995. Support vector networks. *Machine learning* 20:273 297.
- Diaz-Uriarte, Ramon. 2009. VarSelRF: Variable selection using random forests. URL <http://ligarto.org/rdiaz/Software/Software.html>, R package version 0.7-1.
- Erickson, Donna. Articulation of extreme formant patterns for emphasized vowels. *Phonetica* 59:134-149.
- Evgeniou, T., M. Pontil, and T. Poggio. 2000. Regularization networks and support vector machines. *Advances in Computational Mathematics* 13:1–50.
- Fisher, R. A. 1936. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* 7:179–188.
- Fry, D. B. 1958. Experiments in the perception of stress. *Language and Speech* 1:126-152.
- Fry, D. B. 1955. Duration and Intensity as Physical Correlates of Linguistic Stress. *Journal of the Acoustical Society of America* 27:765-768.
- Godfrey, John J., Edward Holliman and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE ICASSP-92. ACL Workshop on Discourse Annotation*.
- Gregory, Michelle L.. & Altun, Yasemin. 2004. Using conditional random fields to predict pitch accent in conversational speech, in ‘Proceedings of ACL’.
- Gussenhoven, Carlos. 1983. Testing the reality of focus domains, *Language and Speech* 26:61–80.

- Heldner, M. 2001. Spectral emphasis as an additional source of information in accent detection, in *Proceedings of Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, pp. 57–60.
- Heldner, M. & Strangert, E. (1998), On the amount and domain of focal lengthening in Swedish two-syllable words, in eds. P. Branderud & H. Traunmüller *Proceedings of FONETIK 98*, Department of Linguistics, Stockholm University, pp. 154–157.
- Howell, Jonathan. 2011. *Meaning and Intonation: On the Web, in the Lab and from the Theorist's Armchair*. Doctoral Dissertation, Cornell University.
- Howell, Jonathan. 2010. Second occurrence focus and the acoustics of prominence. In *Interfaces in Linguistics: New Research Perspectives*, eds. R. Folli & C. Ulbrich. Oxford University Press, 278-298.
- Howell, Jonathan and Mats Rooth. 2009. Web harvest of minimal intonational pairs. In *Proceedings of the Fifth Web as Corpus Workshop*, eds. Iñaki Alegria, Igor Leturia, Serge Sharoff, 45-52. San Sebastian, Spain: Elhuyar Fundazioa.
- Hsu, C. W, C. C Chang, and C. J Lin. 2003. A practical guide to support vector classification. Ms.
- Hirschberg, Julia. 1993. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence* 63:305-340.
- Hyman, Larry. 1978. *Elements of tone, stress and intonation*. Southern California Occasional papers in Linguistics 4.
- Jakobson, Roman, C. Gunnar Fant and Morris Halle 1951. *Preliminaries to Speech Analysis*. Cambridge, MA. MIT Press.
- Joachims, T. 2005. A support vector method for multivariate performance measures, in *Proceedings of the International Conference on Machine Learning*.
- de Jong, Kenneth J. 1995. The supraglottal articulation of prominence in english: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America* 97:491–504.
- Keerthi, S. S, and C. J Lin. 2003. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural computation* 15:1667–1689.
- Heejin Kim & Jennifer Cole. 2005. Acoustic expansion of accented vowels in American English. Presented at the 79th Annual Meeting of Linguistic Society of America, Oakland, CA.
- Klecka, William. 1980. *Discriminant analysis*. Beverly Hills Calif.: Sage Publications.
- Kochanski, Gregory. 2006. Prosody beyond fundamental frequency. In *Methods in Empirical Prosody Research*, eds. S. Sudhoff, D. Lenertová, R. Meyer, S. Pap-

pert, P. Augurzky I. Mleinek, N. Richter, and J. Schließer. Published in Berlin, New York:De Gruyter

Lachenbruch, Peter. 1975. *Discriminant analysis*. New York: Hafner Press.

Ladd, D. Robert. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D. Robert. 1991. One word's strength is another word's weakness: Integrating syntagmatic and paradigmatic aspects of stress. In *Proceedings of the Seventh Eastern States Conference on Linguistics, ESCOL*.

Ladefoged, Peter. 1975. *A Course in Acoustic Phonetics*. New York: Harcourt.

Lehiste, Ilse. 1970. *Suprasegmentals*. Cambridge, MA. MIT Press.

Lieberman, P. 1960. Some acoustic correlates of word Stress in American English. *Journal of the Acoustical Society of America* 32:451-454.

Lin, H.-T. and C.-J. Lin. 2003. A study on sigmoid kernels for SVM and the training of non- PSD kernels by SMO-type methods. Technical report, Department of Computer Science, National Taiwan University.

Mo, Y. 2009. F0 max and formants (F1, F2) as perceptual cues for naïve listeners' prominence perception. Poster presented at the 35th Annual Meeting of the Berkeley Linguistic Society.

Mo, Y. 2008. Acoustic correlates of prosodic prominence for naïve listeners of American English. In *Proceedings of the 34th Meeting of the Berkeley Linguistic Society*. Mukherjee et al. 1999

Pan, S. & McKeown, K. 1999, Word informativeness and automatic pitch accent modeling, in *Proceedings of EMNLP*.

Pierrehumbert, Janet. 1980. *The Phonology and Phonetics of English Intonation*. Cambridge, MA. MIT Press.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org>.

Rooth, Mats. 1996a. On the interface principles for intonational focus. In T. Galloway and J. Spence (eds.) *Proceedings of SALT VI*, 202-226, Ithaca, NY. Cornell University.

Selkirk, Elisabeth. 1995. Sentence prosody: Intonation, stress, and phrasing. In ed. J.A. Goldsmith, *Handbook of Phonological Theory*, Blackwell, 550-569.

Sluijter, A. M. C. and Heuven, V. J. van. 1996a. Acoustic correlates of linguistic stress and accent in Dutch and American English. *Proceedings of ICSLP '96*.

- Sluijter, A. M. C. and Heuven, V. J. van. 1996b. Spectral balance as an acoustic correlate of linguistic stress, *Journal of the Acoustical Society of America* 100:2471-2485.
- Sluijter, A. M. C., Heuven, V. J. van, and Pacilly, J. J. A. 1997. Spectral balance as a cue in the perception of linguistics stress. *Journal of Acoustical Society of America* 101:503 – 513.
- Stevens, J. 2002. *Applied multivariate statistics for the social sciences*. Lawrence Erlbaum.
- Sun, X. 2002 Pitch accent prediction using ensemble machine learning, in *Proceedings of ICSLP 2002*, Denver, Colorado.
- Trubetzkoy, Nikolai S. 1939. Grundzüge der Phonologie. *Travaux du cercle linguistique de Prague* 7. Gottingen: Vandenhoeck and Ruprecht.
- Venables, W. N, and B. D Ripley. 2002. *Modern applied statistics with S*. Springer verlag.