Acoustic Classification of Focus: On the Web and in the Lab

Jonathan Howell[*] Mats Rooth[†] Michael Wagner[‡]

December 11, 2016

Abstract

We present a new methodological approach which combines both naturally-occurring speech "harvested" on the web and speech data elicited in the laboratory. This proof-of-concept study examines the phenomenon of focus sensitivity in English, in which the interpretation of particular grammatical constructions (e.g. the comparative) is sensitive to the location of prosodic prominence. Machine learning algorithms (support vector machines and linear discriminant analysis) and human perception experiments are used to cross-validate the web-harvested and lab-elicited speech. Results confirm the theoretical predictions for location of prominence in comparative clauses and the advantages using both web-harvested and lab-elicited speech. The most robust acoustic classifiers include paradigmatic (i.e. un-normalized), non-intonational acoustic measures (duration and relative formant frequencies from single segments). These acoustic cues are also significant predictors of human listeners' classification, offering new evidence in the debate whether prominence is mainly encoded by pitch or by other cues, and the role that utterance-normalization plays when looking at non-pitch cues such as duration.

1 Introduction

> The World Wide Web is enormous, free, immediately available, and largely linguistic. As we discover, on ever more fronts, that language analysis and generation benefit from big data, so it becomes appealing to use the Web as a data source. The question, then, is how. (Kilgarriff, 2007)

Linguists have been using text data from the web in published work since at least Grefenstette (1999). The appeal, as Kilgarriff (2007) notes, is the low cost

[*] Montclair State University
[†] Cornell University
[‡] McGill University

of entry: most researchers have quick and easy access to a search engine—even quicker since 2007, following the rise of internet-enabled mobile devices.

The cost of entry for *speech* research on the web remains considerably higher, even as the publication of new speech data on the web accelerates, due to platforms like iTunes and YouTube and to infrastructure with greater bandwidth and storage capacity.

The problem, of course, is that search engines search text, not speech. Unless a transcription exists, speech data on the web are effectively invisible. And even when a transcription does exist, it may not be time-indexed. This problem is shared by those who publish speech to the web, and who want to maximize public exposure to their content. For these content producers, there is an obvious commercial incentive to create searchable, time-indexed transcriptions.

At this time, speech researchers do not have access to anything approaching the power and scope of a Google text-search. Nonetheless, the quantity of transcribed, time-aligned speech online is significant and growing. In this study, we present a proof-of-concept for harvesting and analyzing speech data from the web. We leveraged two websites (Everyzing.com and play.it) which indexed radio podcasts with transcriptions obtained with automatic speech recognition (ASR). While the transcriptions varied in quality at the sentence level, accuracy typically exceeded 50% at the level of short, common word sequences (Howell & Rooth, 2009).

Our greatest motivation in using web-harvested speech is the dramatic expansion and diversification of the empirical base it offers linguistic theory. Data collected through personal introspection or elicitation in a university or commercial laboratory represent only a small part of the diversity of human speech. Web-harvested speech, because it is naturally-occurring, reflects a more diverse set of speakers and social contexts. Traditional, curated corpora are static (usually by design) and out-dated (because of the time and effort required to assemble and maintain them). The web, by contrast, is dynamic and evolving.

Naturally, speech data from the web cannot replace existing forms of data. Kilgarriff (2007) cautioned researchers on the challenges and pitfalls of using textual data from the web. In the case of speech data, one must proceed with even greater caution, since the text itself, particularly if generated by ASR, introduces additional biases. With this in mind, we implement an approach for validating results obtained from web data by using machine learning and speech data elicited in the laboratory.

We use our approach to investigate the phenomenon of focus sensitivity in English, in which the interpretation of particular grammatical constructions are sensitive to the location of prosodic prominence. Formal semantic theories of focus make clear predictions for the location of focus in a given discourse context and we want to test these predictions by measuring acoustic prominence in naturally-occurring speech.

Unlike the case of introspective or laboratory data, we lack control over the discourse context in naturally-occurring speech, and it may also be challenging to recover the discourse context. For this proof-of-concept study, then, we investigated a focus-sensitive construction in which the relevant discourse antecedent is always explicit. This is the comparative clause *than I did*, where as we explain below, the location of prominence is predictable from a property of the main clause of the comparative sentence.

We test semantic predictions by building an acoustic classifier. Our methodology uses an explicit machine-learning classification model to evaluate the predictions of a semantic/pragmatic theory of the distribution of focus in a specific syntactic-lexical context, and to investigate the signal features that are involved in marking focus in that context. The methodology is computational and has a certain complexity, but manageable because it takes advantage of well-established classification models and implementations of them in R. In the interest of making it easier to apply the computational method to similar problems, we have distributed code and datasets for the experiments (Howell, 2016).

In addition to evaluating the semantic theory, it is also possible to ask which sets of acoustic measurements contribute to successful classification. We compare the performance of classifiers which use measures of F_0 against classifiers which use other, non-F_0 measures. And we also compare classifiers which use un-normalized, "paradigmatic" measures against classifiers using "syntagmatic" measures which have been normalized within the utterance.

Finally, we also compare the performance of the machine learning classifiers to that of human listeners, and we test whether the same acoustic measures contribute to human listener performance.

The paper has the following organization. The rest of the introduction elaborates the semantics and pragmatics of focus sensitivity and the phonetics and phonology of prosodic prominence. Section 2 describes the methods of data collection for the laboratory-elicited and web-harvested speech, while Section 3 details the machine learning classification, including the algorithms used and how they are evaluated. In Sections 4 and 5 we report on classification using web data and lab data, respectively. In Section 6, we provide a comparison of logistic regression models; and in Section 7, we report on human listener classification. The conclusion is presented in Section 8. Replication data, including acoustic measurements, scripts and speech recordings, to the extent possible, are published online at the Harvard Dataverse (Howell, 2016).

1.1 The semantics of focus sensitivity

In English, prosody is used to mark certain parts of an utterance as salient in the discourse. For instance, the speaker A in (1) makes it salient that someone ate the sushi, but at the time of B's utterance it is not yet salient that Sara ate the sushi. We say that *Sara* is "focused" in (1B) and *ate the sushi* is not focused or "given".

(1) A: You ate the sushi.
B: No, SARA ate the sushi.

In "anaphoric" or "givenness" theories of focus, we understand the relationship between an utterance and the discourse as a kind of anaphora. Roughly, reduced prominence on *ate the sushi* in (1B) is licensed by the earlier sequence *ate the sushi* in (1A).

Discourse anaphors need not be explicit, however. Suppose we are at a Japanese restaurant and order a plate of sushi. You leave for a few minutes and return to find your partner sitting in front of a dirty, empty plate. That someone ate sushi (whether true or not) is now salient in this context, and the utterance in (2) is therefore felicitous, even without an explicit discourse antecedent.

(2) In the presence of a dirty, empty plate...

B: SARA ate the sushi.

Rooth (1992) and Schwarzschild (1999) offer two well known formalisms of focus anaphoricity, the former emphasizing contrastive focus and the latter emphasizing givenness/newness. Both accounts posit a kind of focus skeleton, a semantic object with variables replacing the focused phrases, e.g. ‘X ate the sushi’. Rooth’s “focus semantic value” is the set of propositions obtained by replacing the focused phrase with alternatives of the same type; Schwarzschild achieves a similar effect by existentially quantifying over focused phrases.[1]

Thus, from the utterance $[\text{Sara}]_{\text{F}}$ ate the sushi, Rooth would derive a focus semantic value such as {‘Juan ate the sushi’, ‘The server ate the sushi’, ‘The woman at the next table at the sushi’, ... }. Schwarzschild would derive an existentially quantified proposition ‘Someone ate the sushi’.

Focus is licensed if the focus skeleton stands in a particular relation to a discourse antecedent. For Rooth, the antecedent must be an element of the focus semantic value. For Schwarzschild, the antecedent must entail the existentially quantified proposition. Although more work comparing the two formalisms is required, they are largely equivalent with respect to focus anaphoricity (cf. Rooth 2016). For the following discussion, we’ll assume a relation of entailment holds, following Schwarzschild.

Comparative clauses have the useful property of always occurring with an explicit antecedent: the main clause. Suppose the comparative clause *than I did* in(3 a) is interpreted as ‘I stayed to some degree long’. With focus on the subject *I*, we derive an existentially quantified proposition ‘Someone stayed to some degree long’, which is entailed by the main clause antecedent ‘He stayed to some degree long’.[2]

Similarly, we derive an existentially quantified proposition ‘I like that song some degree at some time’ in (3 b), which is entailed by the main clause antecedent ‘I like that song some degree at the present time’. In (3 c), we derive an existentially quantified proposition ‘I understand some degree little at some time’, which is entailed by the main clause antecedent ‘I understand some degree little today’.

(3) (a) **He** stayed longer than $[\textbf{I}]_{\text{F}}$ did Class “s”
antecedent: He stayed x long

(b) **I** like that song a lot more than **I** $[\text{did}]_{\text{F}}$ Class “ns”
antecedent: I like that song x much

(c) **I** understand less today than **I** did [yesterday]$_F$ Class “ns”
antecedent: I understand x little

As a proxy generalization[3] for focus anaphoricity, we will say that when reference varies in the subject position between the main and *than*-clauses as in (3 a), the subject pronoun *I* in the *than*-clause is semantically focused. When reference is constant in the subject position as in (3 b) and (3 c), semantic focus occurs instead on *did* or on a following adverbial. We can refer to this generalization as the co-reference criterion (4).

(4) **Co-reference criterion for focus in comparative clauses**

If the subjects of the main and comparative clauses have different referents, the token belongs to class “s” (subject focus);

Else, the token belongs to class “ns” (non-subject focus).

With the co-reference criterion, we have an independent way of classifying the comparatives that does not involve prosody.

Together with an interface principle that relates semantic focus to prosodic prominence, theories of focus anaphoricity make testable predictions for the location of prosodic prominence in comparative clauses. A naïve interface principle states simply that a focused constituent is prosodically prominent. One reason for the naïvety of this principle is the non-trivial computation of “focus constituent”. For example, there is a large literature on focus projection (e.g. Gussenhoven 1992; Drubig 1994, 2003; Selkirk 1995; Winkler 1996; Jacobs 1999; Breen et al. 2010) concerned with prominence within large focus constituents (cf. I love [CHEESE]$_F$ vs. I [love CHEESE]$_F$), arguing against a naïve principle. We set aside this issue here, since the focused constituent at issue in our datasets consists of a single element, namely the pronoun *I*.

1.2 The prosody of focus sensitivity

As described in Section 7, we extracted more than 300 acoustic measurements from utterances of *than I did.* In building acoustic classifiers, we do not attempt to make an exhaustive comparison of different combinations of these 309. Nor do these 309 measurements exhaust the possible ways of measuring utterances of this short string. We do, however, consider two ways of grouping the measurements which bear on long-standing issues in the phonetic and phonological study of prosodic prominence.

The first grouping separates syntagmatic measurements (for example, those which relate *I* and *did* in the same utterance) from paradigmatic measurements (for example a measurement from *I* alone). In the last half century, phonologists studying how we produce and perceive prosodic prominence and semanticists studying how we use prosodic prominence to make discourse coherent have advanced their understandings using two ostensibly opposite conceptions of prosodic prominence. Phonologists have argued that prominence should be understood as primarily relational or syntagmatic, e.g. a word or syllable is prominent only with respect to an adjacent word or syllable; semanticists have operated under the tacit assumption that prominence is essentially absolute or paradigmatic: e.g. a word or syllable simply is or isn't prominent.

(5) ... $[\ \alpha\ ]_{\mathrm{F}}\ \beta$ ... paradigmatic comparison
↕
... $\alpha\ \beta$ ...

(6) ... $[\ \alpha\ ]_{\mathrm{F}}\ \beta$ ... syntagmatic comparison
↔

Although linguists from many theoretical traditions have noted the syntagmatic nature of prosody (e.g. Saussure 1967[1916]; Jakobson et al. 1951; Lehiste 1970; Ladefoged 1975), the relational nature of prominence was explicitly codified in the theory of metrical phonology (e.g. Liberman 1975; Liberman & Prince 1977; Hayes 1981; Prince 1983; Selkirk 1984; Halle & Vergnaud 1987; Giegerich 1985), which views prominence, particularly stress, as hierarchically organized rhythmic structure.

In contrast, there is a tradition among semanticists and syntactians to use capitalization, italics or other typographical conventions to indicate prominence, tacitly assuming a paradigmatic comparison. One also finds this view represented in phonetic alphabets, such as the International Phonetic Alphabet, and in early generative theories of prominence (Chomsky & Halle 1968). More recently, several semantic accounts have attempted to model the semantics after the syntagmatic phonological accounts, evaluating focus or givenness as a relation between pairs of adjacent constituents (e.g. Williams 1997; Wagner 2005, 2006; Rooth 2009, 2015).

The second grouping separates measures of F_0 from all other measures. The work of Fry (1955, 1958) long ago dispelled the myth that prominence was realized primarily by loudness. Since then, however, the scientific literature on acoustic prominence has been dominated by discussion of fundamental frequency and pitch. Kochanski (2006) reported that, in one sample, articles about F_0 outnumbered articles investigating other prosodic cues by nearly 5 to 1. Yet different lines of research have pointed to the robustness of non-F_0 measures. Work in laboratory phonetics and phonology has identified non-F_0 cues of accent in speech production (e.g. Ladefoged 1967; Lehiste 1970; De Jong 1991; Campbell & Beckman 1997; Ladefoged & Loeb 2002; Cho 2006) and in the acoustics of speech (e.g. Lehiste 1970; Beckman & Pierrehumbert 1986).

In the domain of phonology, work in the autosegmental tradition (e.g. Liberman 1975; Goldsmith 1976; Bruce 1977; Leben 1973; Pierrehumbert 1980) motivated a distinction between pitch accent and stress. This leads to the question of which category—pitch accent or stress—is the primary correlates semantic focus: in the derviationally-oriented terminology of Selkirk (1984), whether focus is "stress-first" or "accent-first".

Semanticists remained largely unconcerned with the debate, apart from intensive investigation of the licensing configuration for putatively "accentless" second occurrence focus (e.g. Partee 1991; Kadmon 2001; Beaver & Clark 2008; Rooth 1996). Experimental studies of this phenomenon (e.g. Bartels 2004; Beaver et al. 2007; Bishop 2008; Howell 2011) confirmed that significant pitch cues of prominence were indeed absent. However, other acoustic measures of prominence related to stress, such as duration and intensity, were present in small but statistically significant amounts.

Experimental evidence also suggests at least three categorical levels of prominence. Beckman & Edwards (1994) studied the articulation of the syllable *pa* in three contexts, which we will refer to as *phrase accented*, *word accented* and *unaccented*: the first syllable of *papa* (7a); the first syllable of *papa* in (7b) and the second syllable of *papa* in (7b), respectively. The phrase-accented syllable carries a pitch accent and has an unreduced vowel; the prosodic word-accented syllable is postnuclear and has an unreduced vowel; the unaccented syllable has a reduced vowel.

(7) a. [Was her mama a problem about the wedding?]

Her **PA**PA posed a problem.

b. [Did his dad pose a problem as far as their getting married?]

HER **papa** posed a problem.

This categorical distinction was first proposed by Bolinger (1958, 1981) and Vanderslice & Ladefoged (1972) (Gussenhoven 2004:20; see also Halliday (1967)). This influential distinction between phrase accenting and word accenting is fundamental to the ToBI annotation framework (Beckman & Ayers, 1994).

Beckman & Edwards observe that the contrast between the accented syllable and the unaccented syllable is particularly robust for vowel duration and the degree and speed of jaw opening movement. We can infer that vowel reduction is also correlated with less extreme formant movement. Although we do not use them here, measures of spectral balance and post-focal compression have also been implicated in distinguishing between levels of stress (e.g. Sluijter & van Heuven 1996; Xu et al. 2004)

2 Methods of Data Collection

2.1 Web Harvested Data

We collected two different web-harvested corpora of utterances containing “than I did”, using a methodology detailed in Howell & Rooth (2009). A set of standard UNIX tools (e.g. *curl, cutmp3, awk, bash, make*) replicates user interaction with an external search engine. The search engine, provided by RAMP (formerly Everying), uses automatic speech recognition to index speech and identify possible utterances of a word sequence, in our case “than I did”. The first corpus (*web1*) was collected using their search interface at *Everyzing.com*; the second corpus (*web2*) was collected using their search interface at *play.it*. The Everyzing interface searched content from a variety of content providers, but predominantly radio stations, including WEEI, WNYC, KPBS, WRKO, NPR and the White Rose Society. The *play.it* interface searched content from various member stations of CBS radio.[4]

Retrieval efficacy varies by dataset, but Howell & Rooth found that roughly 50% of purported tokens were true, unique and readable. Manual filtering was required. Dataset *web1* contained 90 true tokens of “than I did”: 45 tokens with subject focus (“s”) and 45 tokens with non-subject focus (“ns”). Dataset *web2* contained 127 true tokens: 62 tokens with subject focus and 65 tokens with non-subject focus.

The antecedent and comparative clause in each token was manually transcribed into English prose. From this transcription, the tokens were manually categorized into one of the two focus categories, according to the co-reference criterion (cf. 4). Although this semantic classification was performed by humans, the task did not require special expertise or training beyond identifying and comparing grammatical subjects of the two clauses.

2.2 Laboratory-elicited Data

2.2.1 Stimuli

A total of 16 written stimuli were constructed, modeled after attested examples in the web-harvested corpora. Eight of the stimuli contained an ordinary, first occurrence focus (e.g. 8) and the other eight contained both a first occurrence focus and a repeated, second occurrence focus. The conditions for second occurrence focus were created by contrasting an adjective or verb (e.g. *longer* in 9) or by contrasting a degree modifier (e.g. *lot* in 10).

(8) He saw the situation differently *than I did*. FOF stimulus (subject focus)

(9) You worked harder than I did, FOF stimulus (subject focus)
and you worked longer than I did. SOF stimulus (subject focus)

(10) I think Tom said it a little better than I did. FOF stimulus (subject focus)
In fact, he said it a lot better than I did. SOF stimulus (subject focus)

Among the FOF-only stimuli, half were statements (e.g. 11) and half were wh-questions (e.g. 12).

(11) There were a lot of photographers who would shoot more than I did.

(12) Why do I have more energy today than I did the day before?

Each experimental condition was balanced for semantic focus condition: half of the tokens had subject focus and half had non-subject focus. The full set of stimuli are given in Appendix B. In order to limit the scope of this paper, we leave the SOF examples for future analysis. Henceforth, any mention of laboratory data (*lab*) will refer only to the FOF examples.

2.3 Recording

Participants were recorded in a sound-attenuated room. Twenty-seven individuals participated, although one participant's speech failed to be recorded, leaving a total of 26 participants. Participants were paid.

The stimuli were presented on a computer screen using a set of MATLAB scripts written by Michael Wagner for conducting prosody experiments. In addition to the 16 target sentences, participants also read 18 filler sentences.

No additional context was provided to participants outside of what appears in Appendix B. We choose the comparative construction in order to avoid the challenges of ensuring that each participant used the same discourse antecedent. Since the main clause provides the explicit discourse antecedent for the comparative clause, theory predicts that additional context will be unnecessary for the purpose of conditioning focus.

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, suggesting that the stimuli were reasonably naturalistic.

Nineteen tokens were discarded due to speaker disfluencies, such as false starts, hesitations or utterances that did not match the written stimuli, leaving 397.

2.4 Segmentation

The extraction of acoustic information required annotation at the phonetic level. For each utterance of "than I did", the following phonetic segments were annotated: 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*.

The web-harvested data were labeled manually by the experimenters or by research assistants trained for the task. For segmentation criteria, we used oral and nasal constriction landmarks in the spectrogram and waveform: change in amplitude between vowels and the nasal and oral stops, and the high frequency burst of oral stop releases (cf. Turk et al. 2006).

The laboratory-elicited data were, in addition, automatically forced-aligned using a set of Python scripts that interface with the Hidden Markov Model Toolkit (HTK) (Gorman et al., 2011). Since the manually-annotated laboratory data did not result in improved classification, we report only on the forced-aligned laboratory data. Alignment failed on 3 files for a total of 394 tokens.

2.5 Acoustic Extraction

A total of 309 acoustic measures were extracted using the scripting function of Praat (Boersma & Weenink 2013). Phenomena of interest included duration, fundamental frequency (F_0), first and second formants (F1 & F2), intensity, amplitude, voice quality and spectral tilt. Means or extrema were taken for these phenomena, at regular intervals within a vowel or at the time of other extrema. The ratio between *I* and *did* were also calculated for many measurements, including duration, F_0 and intensity. The full list of measurements is provided with descriptions in Appendix A.

3 Machine Learning Classification

A traditional acoustic phonetic study examines a handful of variables relative to a large number of observations (*small p, large n*), allowing application of methods from classical statistics such as ANOVA and logistic regression. Given a set of more than 300 variables and as few as 90 tokens[5], the data in this study may, by contrast, be considered *large p, small n* (also known as *High Dimension Low Sample Size* or, henceforth, HDLSS). Accordingly, we pursue statistical methods for high-dimensional data borrowed from fields where HDLSS data are ubiquitous, including biotechnology, medical imaging, astrophysics, finance and e-commerce. In genetics, for example, one may wish to examine many thousands of genes in a modest number of tissue samples.

First, we apply two machine learning techniques which have been effective in the classification of HDLSS datasets: linear discriminant analysis (LDA) and support vector machines (SVMs). Second, and in addition, we apply feature selection, using an automated method and a manual, human method. SVMs map linear features into a multidimensional feature space, while feature selection (also known as feature reduction) reduces an apparently high-dimensional structure to a low-dimensional structure.

The machine learning classification and feature selection methods are discussed in Sections 3.1 and 3.2, respectively. Sections 3.3 and 3.4 describe the methods of evaluating classifier performance and the division of data into test sets and training sets.

3.1 Classification Algorithms

Two machine learning techniques were used to create predictive models of the data. Linear discriminant analysis (LDA) is a classification framework based on multidimensional Gaussian probability distributions that has been used widely in pattern recognition tasks (Venables & Ripley, 2002). Support vector machines (SVMs) (Boser et al., 1992) 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). In both cases, we begin with training data consisting of vectors of acoustic measurements, divided into an s set from tokens with shifting reference in the subject position, and ns for constant reference in the subject position. An estimation procedure produces a real-valued objective function h of the linear form (13 a). It can be used used to label points in the space with s or ns, according to the decision rule (13 ab).

The decision surface for the model is the surface that divides points that are classified as s from those classified as ns. In a dataset with two dimensions, the decision surface is a line dividing the two-dimensional space, and in general, in a dataset with n features, a hyperplane (i.e. an affine subspace of dimension $n - 1$). Figures 1 and 2 illustrate decision surfaces in two-dimensional and three-dimensional models drawn from our data.

(13) (a) $h(x) = w \cdot x + b$

(b) if $h(x) > 0$ then s else ns

An LDA model is estimated by fitting a multivariate Gaussian distribution to the data with each label, subject to a constraint of equal co-variance for the two distributions. A Bayes optimal decision rule then results in a linear decision surface. In contrast to an LDA model, which because of the estimation procedure is sensitive to all the training data, the decision plane in an SVM model is sensitive only to a subset of the training data. The plane is positioned in a way that maximizes distance to nearby data points (the support vectors), and includes also a penalty for mis-classified data (Cortes & Vapnik 1995).

LDA models make assumptions which may not be satisfied by the true distributions for our problem, namely normality of the distributions, and (assuming normality) equality of covariances for the two classes. Poor results may also obtain if the training set is small.[6] 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 when the attributes are not intercorrelated (cf. Brown & Wicker (2000)).[7] Our 309 acoustic measurements outnumber the tokens in our datasets, and groups of features are likely to be highly correlated. In the next section, we discuss a method of attribute selection, in order to reduce this number of attributes. In practice, however, it is often possible to obtain good results for an LDA classifier even with small datasets and even with data in violation of the assumptions of normal distribution and homogeneity of covariances (e.g. Lachenbruch 1975; Klecka 1980; Stevens 2002). The implementation of linear discriminant function analysis we use is available in the MASS package (Venables & Ripley 2002) for the statistical computing environment R (R Development Core Team (2013)). The implementation of SVM we use is available in the libsvm package (Chang & Lin 2011; Dimitriadou et al. 2009) for R.[8]

Another feature of SVMs is the possibility of mapping of linear attributes into a multi-dimensional feature space. This is done by replacing dot products by a non-linear kernel function.[9] This greatly reduces the typical computational complexity of training, at the cost of somewhat increased computational complexity during testing.[10] 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 (e.g. duration and energy in our data). Many kernel functions have been used successfully in different classification tasks. Hsu et al. (2003) recommend a radial basis function (RBF)[11], a non-linear mapping which has been shown to also encompass a linear kernel (Keerthi & Lin 2003) and

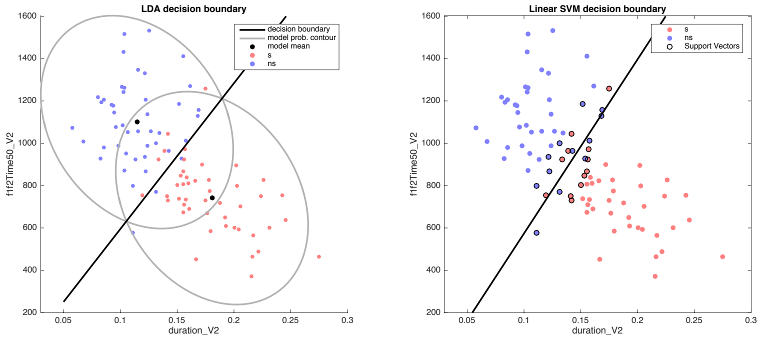

Figure 1: On left, an LDA classier in two dimensions, with ovals marking a contour of equal probability that encloses 75% of probability mass. At right, an SVM classier based on the same training data. Duration-V2 is the duration of the second vowel in *than I did*, and f1f2Time50-V2 is the distance between the first and second formants in the middle of the second vowel. Red points are from observations with varying reference in the subject position (our operational definition of focus), and blue points are from observations with constant reference in subject position. In this case the LDA and SVM decision surfaces are are nearly the same.

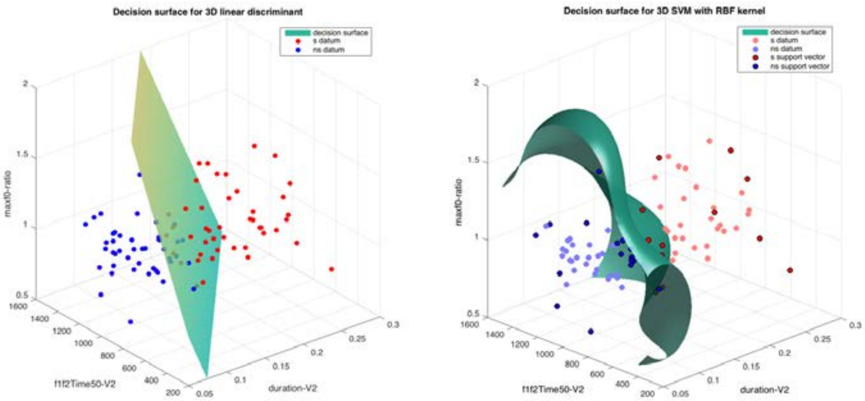

Figure 2: On the left, the decision surface for a three-dimensional LDA classifier. The extra feature relative to Figure 1 is maxf0-ratio, the between the maximal F_0 value in the second vowel of *than I did* and the maximal F_0 in the third vowel. Red points have varying reference in the subject position (subject focus), and blue points are from observations with constant reference in subject position. On the right, the decision surface for an SVM with radial basis function kernel estimated from the same data.

behaves 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, which sets the penalization for a datum occurring on the wrong side of the margin.) 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. This situation obtains for those of our classifiers which use the unfiltered set of 309 attributes and are applied to datasets of 90 and 127 (viz. the web-harvested datasets). We therefore consider classifiers using both RBF kernels in addition to linear ones. Figure 2 shows the curved decision boundary obtained in a three-dimensional model with an RBF kernel.

Estimating LDA and SVM classifiers require datasets without missing values. Algorithms in Praat and other acoustic analysis software have a notoriously difficult time extracting values such as F_0 in the absence of regular, periodic voicing. A dataset with missing values was therefore unavoidable, and many values were undefined. Typically, a dataset with less than 5% missing data is considered manageable, while more than 5-10% missing data may bias subsequent statistical analysis and can require sophisticated methods of data imputation. All of our datasets fell within these acceptable rates of missingness. The web-harvested datasets web1 and web2 had missing rates of 0.6% and 0.8%, respectively. The laboratory-elicited dataset had a missing rate of 2.5%.

Many different kinds of data imputation exist. Single imputation methods replace all missing data with the same value, such as -1, 0 or 1, or the mean or median of the variable. One disadvantage of single imputation is that it fails to model the variability of the underlying data. Multiple imputation methods use algorithms to impute a particular value for each data point using information from observations without missing data. Some common methods of multiple imputation include hot- and cold-deck imputation and k-nearest neighbor imputation. We experimented informally with several of these methods of imputation and none resulted in noticeable differences in classifier performance. Leaving in-depth study of imputation for future research, we chose to use mean imputation on all of the datasets.

Optimizing the value of hyperparameters is often recommended. In this study, we were able to achieve robustly performing classifiers with the default settings ($\gamma = \frac{1}{n}$ and $C = 1$). We therefore leave the contribution of tuning to future investigation.

3.2 Redundant features and feature selection

In building a classifier, one may be concerned simply with the classification task itself: developing and improving the ability of a particular decision function to generalize from one set of data (a training set) to another (a test set). We may call this the "functional measure" (cf. Cristianini & Shawe-Taylor 2000).

One may also be concerned with how the classification task is achieved and how closely it models real human cognitive ability. We may call this the "descriptional measure". The relative importance of the functional and descriptional measures typically varies according to the goals of the researcher. Consider the following functionally-oriented view from Cristianinni & Shawe-Taylor.

> Shifting our goal to generalisation removes the need to view our hypothesis as a correct representation of the true function. [...] In this sense the criterion places no constraints on the size or on the 'meaning' of the hypothesis – for the time being these can be considered to be arbitrary. (Cristianini & Shawe-Taylor 2000: Section 1.2)

Another more descriptionally-oriented researcher concerned primarily with the underlying or "true" function may be wary of even a high-accuracy decision function which incorporates what may seem to be linguistically irrelevant or orthogonal noise in the data.

In practice, however, the functional and descriptional are not mutually exclusive and are, one hopes, mutually informative. One may, for example, apply the functional measure to establish a pattern in the data and apply other methods to understand the contribution of different features in the model. In this study, we want a classifier which accurately predicts a focus category—a functional measure, but we also wish to know which acoustic measures are important for this task—a descriptional measure.

Feature selection is one means of peering into the "black box" of a classifier, and understanding which features are contributing to a model's generalization accuracy. Pragmatically, feature selection is also sometimes necessary to improve classifier performance. Collinearity in LDA models have been shown to lead to stability problems (e.g. Naes 2001). SVMs, despite their promise as a classifier which does not require feature selection, have been shown to improve the generalization accuracy and/or model complexity (and thus computation) for those datasets with redundant and/or irrelevant features. For example, Sarojini et al. (2009) demonstrate improved accuracy for a clinical dataset with a large number of instances (768) and a small number of features (8 prior to feature elimination) while conversely Weston et al. (2001) demonstrate this effect for cancer discrimination in a dataset with a small number of instances (72) and a large number of features (7129 genes). More generally, removing redundant features mitigates the potential for a classifier to be mislead and for overfitting of the model.

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:

> Feature selection should be viewed as a part of the learning process itself, and should be automated as much as possible. On the other hand, it is a somewhat arbitrary step, which reflects our prior expectations on the underlying target function. (Cristianini & Shawe-Taylor 2000: Chapter 3)

> To start, the initial variable list should be logically screened, based on substantive theory, prior research, and reliability of measures, as well as on practical grounds. Next, the list can be statistically screened. (Huberty 2006:11)
>
> Of course, an investigator's professional opinion also can be relied upon when selecting potential discriminator variables. (Brown & Wicker 2000:212)

Many statistical methods of feature selection exist. Filter methods select features according to some importance measure independent of the classifier, such as correlation or information gain. Embedded methods incorporate selection into the training process of classification; a set of *minimally optimal* features for the classification task are identified. As a reviewer notes, features which are pre-selected automatically may however cause an increase in the generalization error rate of a classifier (see for example Barron 1994). In contrast, wrapper methods use information from a classifier (possibly a different classifier) prior to training and are used to select not just a set of non-redundant features, but *all relevant* features. With a functional measure in mind, we chose an all-relevant wrapper method known as the Boruta algorithm and available as an R package (Kursa & Rudnicki, 2010). Briefly, the algorithm generates fake or "shadow" features and iteratively compares the real features against them, using a random forest classifier to compute a significance measure.

In addition to applying the Boruta algorithm to the full set of 309 acoustic measures, we also applied the algorithm to theoretically meaningful subsets: F_0-related measures, non-F_0-related measures, syntagmatic measures (i.e. ratios between *I* and *did*), and paradigmatic measures (i.e. from a single word). Finally, based on a combination of theoretical expectation and trial-and-error, we also selected several feature sets by hand.

Note that because we are considering two different training sets (first the web-harvested dataset *web1* and later the laboratory-elicited dataset *lab*), we apply feature selection independently for the two sets. The results of the Boruta algorithm are detailed in Section 4.1.

The set of acoustic measures used by a machine learning classifier to predict focus will not necessarily correspond to the set of acoustic measures that a human speaker uses to convey focus or to the set of features an individual human listener uses to interpret focus. It is therefore important to investigate the use of acoustic measures in human classification, which we do in Section 7.

3.3 Evaluation of classifier performance

Typically, a classification algorithm generates a model from a set of labeled data (a training set) and this model is then used to predict unseen data without labels (a test set). If the correct labels of the test set are known, we can compare them against the model's predictions. The proportion of correct labels and the proportion of incorrect labels are known as the generalization accuracy and generalization error, respectively. As a measure of bias–whether the classifier tends to predict one class more accurately than the other–we calculated a balanced error rate, which is an average of the two within-class error rates. We also compare the results against a simple baseline accuracy, which is the proportion of the larger class. The three statistics are given in (14,15,16).

(14) Baseline accuracy $\dfrac{\#\text{tokens in largest class of test set}}{\#\text{ tokens in both classes in test set}}$

(15) Generalization accuracy $\dfrac{\#\text{tokens in test set accurately classified}}{\#\text{ tokens in test set}}$

(16) Balanced error rate $\left(\dfrac{\#\text{tokens incorrect``s"}}{\#\text{ total``s"}} + \dfrac{\#\text{ incorrect``ns"}}{\#\text{ total``ns"}}\right) \cdot \dfrac{1}{2} \cdot 100$

In addition to calculating these performance statistics for each classifier, we also wanted some confidence that a classification model wasn't overfitting the particular training data and that its performance on the test data was not by chance. To assess this, we performed permutation-based validation (cf. Hsing et al. 2003; Jensen 1992; Molinaro et al. 2005). The class labels of the training set were randomly permuted before training and performance statistics calculated in the usual way. This process was iterated n times. In theory, one may repeat this for all possible permutations, although this strategy is impractical for computational reasons. A large number of iterations (we chose $n = 5000$) produces a reasonable approximation of the empirical cumulative distribution for the permutation-achieved performance statistics.

In Figure 3, we plot the empirical distribution of permutation-achieved generalization accuracy for a particular classifier. The x-axis represents generalization accuracy; the y-axis represents the cumulative distribution (i.e. the proportion of the permuted data which is less than or equal to the value of x). If the observed accuracy or balanced error falls outside of the 95th or 99th percentile, we say that the observation is statistically significant with a p-value of greater than 0.05 or 0.01, respectively. This provides a confidence measure with which we can reject the null hypothesis that the classifier achieved the observed statistic at random.

We can visualize the significance of multiple classifiers in one figure by plotting the observed and permutation-achieved statistics as single points, as in Figure 4 (cf. Lyons-Weiler et al. 2005). The observed statistic is plotted in black; the permutation achieved statistic at p=0.05 and p=0.01 is plotted in green and red, respectively. If the observed statistic falls outside of the permutation achieved statistic, we say that the observed statistic is statistically significant.

For convenience, we plot the accuracy rate and balanced error rate on the figure. An asymmetry between the two, therefore, represents a bias towards one of the two classes.

More extreme permutation-achieved statistics (i.e. greater accuracy or smaller balanced error) reflect more structure in the data. For example, if the

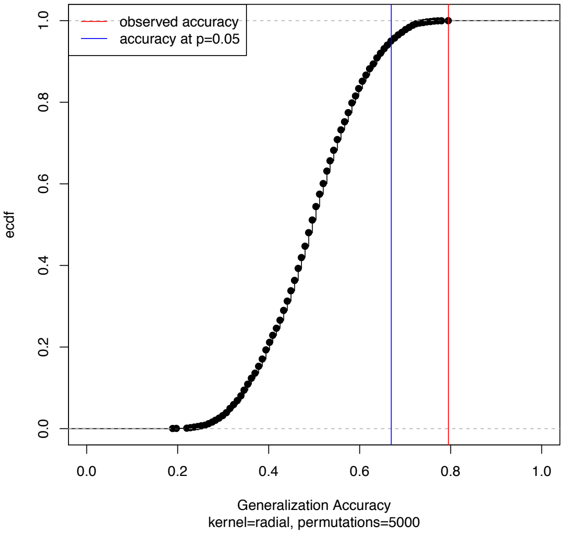

Figure 3: Example plot of an empirical cumulative distribution of permutation-achieved statistics. The x-axis represents generalization accuracy; the y-axis represents the *cumulative* distribution (i.e. the proportion of the permuted data which is less than or equal to the value of x). If the observed accuracy (red line) falls outside of the 95th percentile (blue line), we say that the observation is statistically significant with a p-value of greater than 0.05. This provides a confidence measure with which we can reject the null hypothesis that the classifier achieved the observed statistic at random.

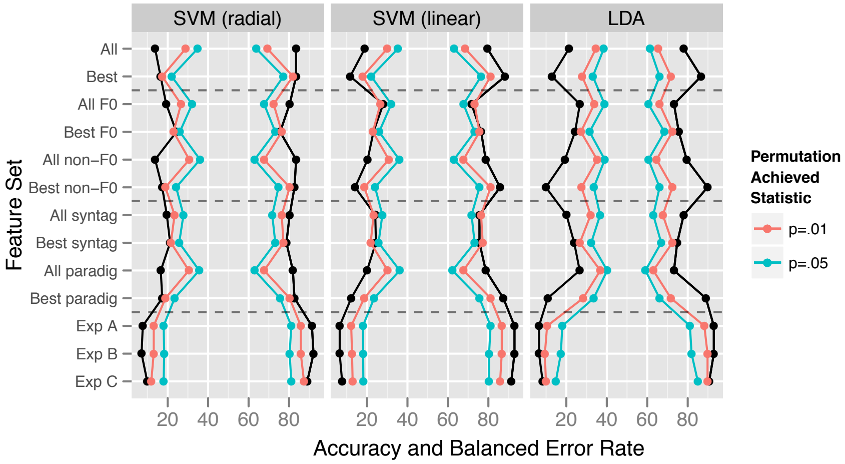

Figure 4: Example plot of permutation-achieved statistics (cf. Lyons-Weiler et al. 2005). Each panel shows a different classification method: SVM with radial kernel, SVM with linear kernal and LDA (left-to-right). Within a panel, the x-axis displays accuracy/error rate as a percentage and the y-axis lists classifiers according to the feature sets used. Within a panel, the left-side black dot corresponds to the observed balanced error rate, while the right-side black dot corresponds to the observed accuracy rate. An asymmetry between the two dots indicates a bias towards one of the two classes. The permutation achieved statistic at p=0.05 and p=0.01 is plotted in green and red, respectively. More extreme permutation-achieved statistics (i.e. greater accuracy or smaller balanced error) reflect more structure in the data. If the observed statistic (black dot) falls outside of the permutation achieved statistic (i.e. to the left of the colored dots in the case of balanced error rate or to the right of the colored dots in the case of accuracy rate), we say that the observed statistic is statistically significant.

Table 1: Summary of Datasets

Name	source	annotation	size (ns/s)	baseline accuracy
web1	web-harvested:*Everyzing*	manual	90 (45/45)	50.5
web2	web-harvested:*play.it*	manual	127 (65/62)	51.2
lab	laboratory-elicited	automated	394 (193/201)	51

acoustic values are randomly distributed with respect to the two focus classes, we expect less variance among the permutation-based statistics and less extreme statistics at p=.01 and p=.05. If the acoustic values are nonrandomly distributed with respect to the classes, we expect more variance and more extreme statistics at p.=01 and p=.05, since some permutations will be more positively and negatively correlated with the acoustic values.

Note that this permutation test does not directly compare one classifier's performance against another's. In order to determine this, we performed McNemar's test (McNemar 1947; see also Gillick & Cox 1989), a test for the difference of two proportions which has been used to compare machine learning classifiers. A non-parametric counterpart of the t-test, McNemar's test compares the null hypothesis that the two classifiers misclassify the same tokens, using a $\chi 2$ test for goodness-of-fit. Although it doesn't measure variability due to choice of training set (e.g. web1 vs. lab), McNemar's test does exhibit acceptably low Type I error and reasonably high power Dietterich (1998). We apply the test to compare a classifier using a subset of features to the related algorithm using the full set of features.

Finally, we wanted to assess the relevance of individual features used in the best performing classifiers. To do this, we compared pairs of logistic regression models–one with and one without the feature of interest–using an ANOVA and chi-squared test of statistical significance.

3.4 Training/test pairs

In order to make comparison manageable, we did not test and train the datasets in all possible combinations. Rather, we used just two datasets for training: the web-harvested dataset *web1* and the laboratory-elicited dataset *lab*. We tested the web-trained classifiers on the remaining web-harvested dataset *web2* and on the laboratory-elicited dataset *lab* (Section 4). We tested the laboratory-trained classifiers on the web-harvested dataset *web2* (Section 5). Table 1 summarizes the datasets under consideration.

4 Machine Classification Experiments 1: Web-harvested Training Data

4.1 Feature Selection by Algorithm

All-relevant feature selection using the Boruta algorithm applied to the *web1* web-harvested dataset produced the feature sets (17,18,19,20,21). From the full feature set, the algorithm selected a combination of F_0, *non-*F_0, syntagmatic and paradigmatic features that included measures of vowel duration, F_0, energy and formant values. No measures of intensity, spectral tilt, jitter or shimmer were selected.

From the set of exclusively F_0 features, Boruta selected measures of the value and timing of F_0 means, minima and maxima, both paradigmatically and syntagmatically.

From the set of exclusively *non-*F_0 features, Boruta selected measures of vowel duration, glottal pulse, intensity, energy, amplitude and formant values. The duration and formant values were mostly paradigmatic, coming almost exclusively from *I*; and the values related to loudness (i.e. glottal pulse, intensity, energy and amplitude) were all syntagmatic. The formant values came from the first and second formant of *I*, both individually and as a ratio, and at intervals from 20 to 70 percent of the vowel duration.

From the set of exclusively paradigmatic features, Boruta selected measures of duration, glottal pulse, minimum and range of F_0, and first and second formant values at several different intervals. All of the paradigmatic features were selected from *I*.

From the set of exclusively syntagmatic features, Boruta selected measures of intensity, amplitude, energy, duration and the value or timing of mean, minimum, maximum and range of F_0.

(17) Features selected by Boruta algorithm from full feature set

duration_V2
pulses_V2
pulses_ratio
meanf0_ratio
maxf0_ratio
minf0Time_ratio
rangef0_V2
rangef0_ratio
energy_ratio
f2Time20_V2
f2Time30_V2
f1Time40_V2
f2Time40_V2
f1Time50_V2
f2Time50_V2
f1Time60_V2
f2Time60_V2
f1Time70_V2
f1f2Time20_V2
f1f2Time30_V2
f1f2Time40_V2
f1f2Time50_V2
f1f2Time60_V2

(18) Features selected by Boruta algorithm from set of F_0 features

meanf0_ratio
maxf0_ratio
minf0_ratio
maxf0Time_V2
maxf0Time_V3
maxf0Time_ratio
minf0Time_V2
minf0Time_ratio
rangef0_V2
rangef0_ratio

(19) Features selected by Boruta algorithm from set of non-F_0 features

duration_V2
pulses_V2
pulses_ratio
maxIntensity_ratio
energy_ratio
amp_ratio
maxf1_V2
f1Time20_V2
f2Time20_V2
f2Time30_V2
f1Time40_V2
f2Time40_V2
f1Time50_V2
f2Time50_V2
f1Time60_V2
f2Time60_V2
f1Time70_V2
f1f2Time20_V2
f1f2Time30_V2
f1f2Time40_V2
f1f2Time50_V2
f1f2Time60_V2

(20) Features selected by Boruta algorithm from set of syntagmatic features

pulses_ratio
meanf0_ratio
maxf0_ratio
minf0_ratio
maxf0Time_ratio
minf0Time_ratio
rangef0_ratio
maxIntensity_ratio
energy_ratio
amp_ratio
duration_ratio

(21) Features selected by Boruta algorithm from set of paradigmatic features

duration_V2
pulses_V2
minf0Time_V2
rangef0_V2
f1f2Time10_V2
f2Time20_V2
f2Time30_V2
f1Time40_V2
f1Time40_V2
f1Time50_V2
f1Time50_V2
f1Time60_V2
f1Time60_V2
f1Time70_V2
f1f2Time20_V2
f1f2Time30_V2
f1f2Time40_V2
f1f2Time50_V2
f1f2Time60_V2

4.2 Classifier Results

Web-trained, web-tested

In this section we train the classifiers on the web-harvested dataset *web1*, and test the classifiers on a second web-harvested dataset *web2*. The results are summarized in Table 2 and Figure 5. Each of the observed classifier accuracies and balanced error rates of the classifiers was statistically significant (p<.05 using a permutation-achieved statistic), with the best performing classifier achieving 92.9% accuracy and 6.5% balanced error.

The classifiers using the full set of 309 features performed well above the baseline (83.5%, 79.5% and 78.0% with SVM-RBF, SVM-linear and LDA, respectively). Recall, however, that we use feature selection with the intent of removing redundant, potentially misleading features and avoiding overfitting of the model. Most of the classifiers using an automatically selected subset of features failed to achieve a statistically significant improvement (p<.05 using McNemar's test) over the same algorithm using the full set of features, with the exception of the LDA algorithm using a set of best non-F_0 features or a set of best paradigmatic features. In addition, we observed the following numerical tendencies: classifiers using only non-F_0 measures outperformed classifiers using only F_0 measures; and classifiers using only paradigmatic measures outperformed classifiers only using syntagmatic measures.

Table 2: Accuracy and balanced error rates for different classification models: training set *web1*; test set *web2*. Shading indicates p <.05 according to the corresponding permutation-achieved statistic: i.e. rejection of the null hypothesis that the classifier achieved the accuracy/BER by chance. Bolding indicates p<.05 in a comparison with the same algorithm trained on the full feature set, using McNemar's test: i.e. rejection of the null hypothesis that the difference between the two classifiers is due to chance.

	web1 → web2						
Feature set	Baseline	SVM (RBF)		SVM (linear)		LDA	
1. Full feature set	51.2	83.5	13.8	79.5	18.9	78.0	21.3
2. "Best" features	51.2	83.5	16.5	88.2	11.7	86.6	12.9
3. All F_0 features	51.2	80.3	19.3	71.7	28.0	73.2	26.7
4. "Best" F_0 features	51.2	75.6	24.3	76.4	23.2	75.6	24.3
5. All non-F_0 features	51.2	83.5	13.8	78.7	20.3	79.5	19.3
6. "Best" non-F_0 features	51.2	82.7	17.3	85.8	14.1	**89.9**	**9.9**
7. All syntagmatic features	51.2	80.3	19.6	75.6	24.1	78.0	20.1
8. "Best" syntagmatic features	51.2	78.7	21.2	75.6	24.4	74.8	23.9
9. All paradigmatic features	51.2	81.9	16.6	78.7	20.0	73.2	26.6
10. "Best" paradigmatic	51.2	82.7	17.3	87.4	12.2	**89.0**	**10.8**
11. Experimenter-selected A *duration_ V2, f1f2Time50_ V2,* *meanf0_ ratio,* *duration_ C3*	51.2	**91.3**	**7.7**	**92.9**	**6.5**	**92.9**	**6.5**
12. Experimenter-selected B *duration_ V2, f1f2Time50_ V2,* *maxf0_ ratio, duration_ C3*	51.2	**92.1**	**7.1**	**92.9**	**6.5**	**92.9**	**6.5**
13. Experimenter-selected C *duration_ V2, f1f2Time50_ V2,* *duration_ C3*	51.2	89.0	9.9	**91.3**	**7.7**	**90.6**	**8.3**

The different algorithms performed competitively, and the classifiers using experimenter-selected features achieved the best overall results, almost all showing a statistically significant improvement over the same algorithm using the full set of features.

Web-trained, lab-tested

Next, we use the same set of web-trained classifiers and test them on laboratory data. The results are summarized in Table 3 and Figure 6.

The observed accuracies and balanced error rates of nearly all of the classifiers were statistically significant (p<.05 using a permutation-achieved statistic), with the best-performing classifier achieving 87.6% accuracy and 10.5% error. The non-significant results came principally from classifiers using only F_0 features and only syntagmatic features.

The classifiers using the full set of 309 features performed well above the baseline (77.9%, 80.2% and 72.8% with SVM-RBF, SVM-linear and LDA, respectively). Classifiers using only the automatically selected set of best F_0 features demonstrated a statistically significant worse performance (p<.05 using

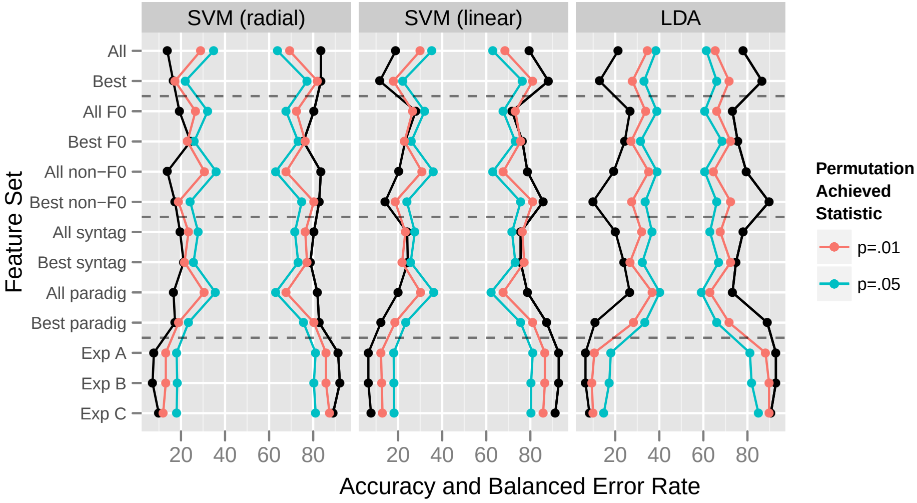

Figure 5: Accuracy and balanced error rates with permutation achieved significance for different classification models: training set *web1*; test set *web2*.

McNemar's test) over the same algorithm using the full set of features. By contrast, classifiers using only the automatically selected set of best non-F_0features achieved a statistically significant improvement. Similarly, classifiers using only the automatically selected set of syntagmatic features failed to achieve a statistically significant improvement or demonstrated a statistically significant worse performance over the same algorithm using the full set of features, while classifiers using only the automatically selected set of best paradigmatic features achieved a statistically significant improvement.

The different algorithms performed competitively, and the classifiers using experimenter-selected features achieved the best overall results, all showing a statistically significant improvement over the same algorithm using the full set of features.

4.3 Discussion

The performance of classifiers trained on web-harvested data overwhelmingly supports the theoretical predictions for location of prominence in comparative clauses. With few exceptions, classifiers performed well above the baseline, calculated as the percentage of the larger class, and satisfied statistical significance ($p<.05$), calculated using permutation achieved statistics.

Variation in classifier performance revealed not only the robustness of classifiers using F_0 and syntagmatic features, but the robustness of classifiers making

Table 3: Accuracy and balanced error rates for different classification models: training set *web1*; test set *lab*. Shading indicates p <.05 according to the corresponding permutation-achieved statistic: i.e. rejection of the null hypothesis that the classifier achieved the accuracy/BER by chance. Bolding indicates p<.05 in a comparison with the same algorithm trained on the full feature set, using McNemar's test: i.e. rejection of the null hypothesis that the difference between the two classifiers is due to chance.

	web1 → lab						
Feature set	Baseline	SVM (RBF)		SVM (linear)		LDA	
1. Full feature set	51.0	77.9	16.9	80.2	16.6	72.8	23.6
2. "Best" features	51.0	85.0	14.2	**85.8**	**12.1**	**85.8**	**13.1**
3. All F_0 features	51.0	**69.0**	**30.5**	**66.5**	**33.5**	70.8	29.2
4. "Best" F_0 features	51.0	**69.0**	**30.6**	**63.2**	**36.6**	**62.4**	**37.5**
5. All non-F_0 features	51.0	79.7	15.3	81.0	15.7	**79.7**	**18.8**
6. "Best" non-F_0 features	51.0	**83.8**	**14.4**	**83.0**	**14.4**	**81.2**	**15.2**
7. All syntagmatic features	51.0	75.4	22.3	**70.8**	**28.6**	**64.5**	**35.5**
8. "Best" syntagmatic features	51.0	73.4	25.4	**67.3**	**32.2**	**64.5**	**35.4**
9. All paradigmatic features	51.0	78.7	16.2	**73.4**	**20.1**	**78.9**	**19.7**
10. "Best" paradigmatic	51.0	76.9	17.8	80.5	15.2	**81.7**	**16.0**
11. Experimenter-selected A *duration_ V2, f1f2Time50_ V2,* *meanf0_ ratio, duration_ C3*	51.0	**84.5**	**14.8**	**85.3**	**13.6**	**85.3**	**13.3**
12. Experimenter-selected B *duration_ V2, f1f2Time50_ V2,* *maxf0_ ratio, duration_ C3*	51.0	**87.3**	**12.0**	**85.3**	**12.8**	**86.3**	**11.6**
13. Experimenter-selected C *duration_ V2, f1f2Time50_ V2,* *duration_ C3*	51.0	**87.6**	10.5	**87.6**	**10.7**	**85.0**	**12.2**

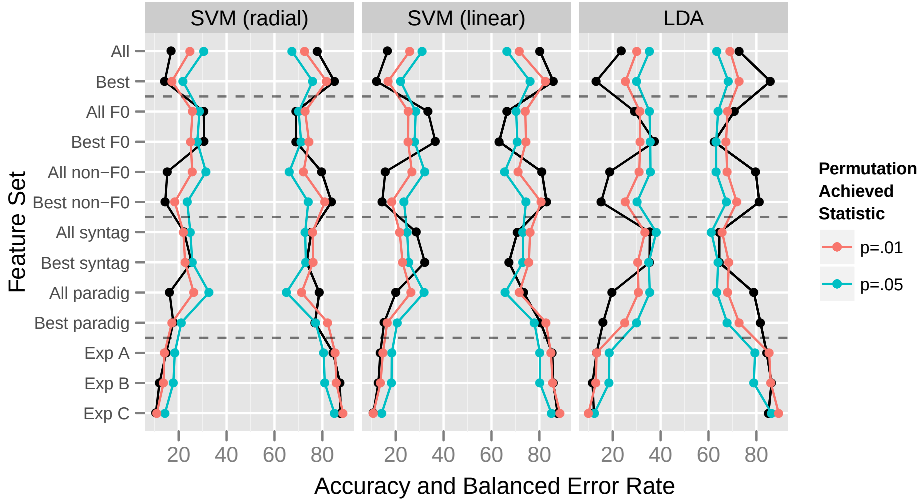

Figure 6: Accuracy and balanced error rates with permutation achieved significance for different classification models: training set *web1*; test set *lab*.

use of non-F_0 and paradigmatic features. Indeed, a majority of classifiers which used exclusively non-F_0 or paradigmatic features achieved statistical significance ($p<.05$), even on those datasets for which many classifiers using exclusively F_0 or syntagmatic features failed to meet statistical significance. It was also the case that neither the classifiers using exclusively F_0 features nor classifiers using exclusively syntagmatic features achieved a statistically significant improvement over the same algorithm using all 309 features. By contrast, classifiers using exclusively non-F_0 or paradigmatic features in most cases achieve a performance which is either not significantly different from or significantly better than the same algorithm using all 309 features.

For each dataset, the best performing classifiers used a combination of F_0, non-F_0, paradigmatic and syntagmatic features, usually those selected by hand: duration of *I*, the first consonant closure duration in *did*, the F1-F2 differential at the midpoint of *I*, and the mean or maximum F_0 ratio between *I* and *did*. Further, the hand-selected classifiers that lacked the latter, syntagmatic F_0 measures did not differ significantly from those which included them (at the level of $p<0.05$ using McNemar's test).

Thus, while the syntagmatic and F_0 measures are undeniably *relevant* to the categorization of focus placement in these data (demonstrated also by their selection by the all-relevant Boruta algorithm), the evidence suggests that they may not be *necessary.* The results of these classification experiments are incompatible with theories of prominence and focus realization which privilege F_0 measures and syntagmatic evaluation to the exclusion of other non-F_0 and paradigmatically evaluated measures.

We also wish to acknowledge that duration may properly be regarded as syntagmatic, given the effect of speech rate on its interpretation. Among our full set of features we included a syntagmatic measure of duration: the ratio of duration of *I* to the duration of *did*. However, unlike the un-normalized measure of duration on *Í* alone, the syntagmatic measure of duration was not selected by the Boruta feature selection algorithm. Anecdotally, the syntagmatic measure of duration was also unhelpful in arriving at an Experimenter-selected set of best features.

5 Machine Classification Experiments 2: Laboratory-Elicited Training Data

5.1 Feature Selection by Algorithm

All-relevant feature selection using the Boruta algorithm applied to the *lab* laboratory-elicited dataset produced much larger feature sets than when applied to the *web1* web-harvested dataset. The feature sets selected are listed in Appendix C. From the full feature set, the algorithm selected 66 different features, which included measures of F_0, glottal pulse, jitter, shimmer, intensity, energy, power, first and second formants and duration, and a combination of syntagmatic and paradigmatic measures.

From the set of exclusively F_0 features, Boruta selected measures of the value and timing of F_0 means, minima, maxima and range, both paradigmatically and syntagmatically.

From the set of exclusively non-F_0 features, Boruta selected 60, which included measures of vowel duration, glottal pulse, jitter, shimmer, intensity, energy, power amplitude and formant values. Both syntagmatic and paradigmatic values were selected, the latter coming from both *I* and *did.* The formant values came from the first and second formant of *I* at intervals from 10 to 80 percent of the vowel duration.

From the set of exclusively syntagmatic features, Boruta selected 15 features, which included measures of amplitude, intensity, energy, power, duration, glottal pulse and F_0.

From the set of exclusively paradigmatic features, Boruta selected 60 features, which measures of duration, glottal pulse, intensity, power, jitter, shimmer, F_0, and first and second formant values, from both *I* and *did.*

5.2 Classifier Results

Lab-trained, web-tested

In this section, we train classifiers on laboratory data (lab) and test them on web-harvested data (*web2* in order to compare results from Section 4). The results are summarized in Table 4 and Figure 7.

The observed accuracies and balanced error rates of nearly all of the classifiers were statistically significant ($p<0.05$), with the best-performing classifier achieving 92.1% accuracy and 7.9% error. The non-significant results came from classifiers using only F_0 features and only syntagmatic features.

The classifiers using the full set of 309 features performed well above the baseline (77.2%, 77.2% and 72.4% with SVM-RBF, SVM-linear and LDA, respectively). Classifiers using only the automatically selected set of best F_0 features failed to achieve a statistically significant improvement ($p<.05$ using McNemar's test) or demonstrated a statistically significant worse performance over the same algorithm using the full set of features. By contrast, classifiers using only the automatically selected set of best non-F_0 features achieved a numerical but not statistically significant improvement.

In addition, we observed the following numerical tendencies: classifiers using only non-F_0 measures outperformed classifiers using only F_0 measures; and classifiers using only paradigmatic measures outperformed classifiers using only syntagmatic measures. Similarly, classifiers using only the automatically selected set of syntagmatic features failed to achieve a statistically significant improvement or demonstrated a statistically significant worse performance over the same algorithm using the full set of features, while classifiers using only the automatically selected set of best paradigmatic features achieved a numerical but not statistically significant improvement.

The different algorithms performed competitively, and the classifiers using experimenter-selected features achieved the best results overall, all showing a

Table 4: Accuracy and balanced error rates for different classification models: training set *lab*; test set *web2*. Shading indicates $p < .05$ according to the corresponding permutation-achieved statistic: i.e. rejection of the null hypothesis that the classifier achieved the accuracy/BER by chance. Bolding indicates $p < .05$ in a comparison with the same algorithm trained on the full feature set, using McNemar's test: i.e. rejection of the null hypothesis that the difference between the two classifiers is due to chance.

	lab → web2						
Feature set	Baseline	SVM (RBF)		SVM (linear)		LDA	
1. Full feature set	51.2	77.2	22.3	77.2	22.7	72.4	27.3
2. "Best" features	51.2	78.0	19.7	78.0	22.0	77.2	22.5
3. All F_0 features	51.2	67.7	32.2	**65.4**	**34.7**	**60.6**	**38.8**
4. "Best" F_0 features	51.2	70.1	29.9	**66.1**	**33.9**	**59.8**	**39.6**
5. All non-F_0 features	51.2	78.7	21.2	74.8	25.0	68.5	31.4
6. "Best" non-F_0 features	51.2	78.0	20.2	78.0	21.4	81.1	18.2
7. All syntagmatic features	51.2	67.7	29.1	**61.4**	**34.7**	63.0	32.5
8. "Best" syntagmatic features	51.2	68.5	28.5	**58.3**	**37.3**	61.4	34.0
9. All paradigmatic features	51.2	78.0	22.0	78.7	20.9	74.0	25.7
10. "Best" paradigmatic	51.2	79.5	19.0	78.0	21.1	74.8	25.0
11. Experimenter-selected A *duration_ V2, f1f2Time50_ V2,* *meanf0_ ratio, duration_ C3*	51.2	**88.2**	**11.6**	**90.6**	**9.2**	**90.6**	**9.2**
12. Experimenter-selected B *duration_ V2, f1f2Time50_ V2,* *maxf0_ ratio, duration_ C3*	51.2	**88.2**	**11.6**	**92.1**	**7.9**	**90.6**	**9.2**
13. Experimenter-selected C *duration_ V2, f1f2Time50_ V2,* *duration_ C3*	51.2	**86.6**	**13.2**	**90.6**	**9.2**	**89.8**	**10.0**

statistically significant improvement over the same algorithm using the full set of features.

5.3 Discussion

The performance of classifiers trained on laboratory-elicited data, like classifiers trained on web-harvested data, strongly supports the theoretical prediction for location of prominence in comparative clauses. With few exceptions, classifiers performed well above the baseline, and satisfied statistical significance ($p<.05$).

The classifiers trained on laboratory-elicited data also revealed the robustness of classifiers using non-F_0 and syntagmatic features. Indeed, classifiers using non-F_0 and syntagmatic features achieved statistical significance without exception. Numerically, the classifiers using non-F_0 and syntagmatic features also outperformed the classifiers using F_0 and paradigmatic features.

The best performing classifiers, however, included a combination of F_0, non-F_0, syntagmatic and paradigmatic, whether the features were selected algorithmically with Boruta or whether the features selected manually by the experi-

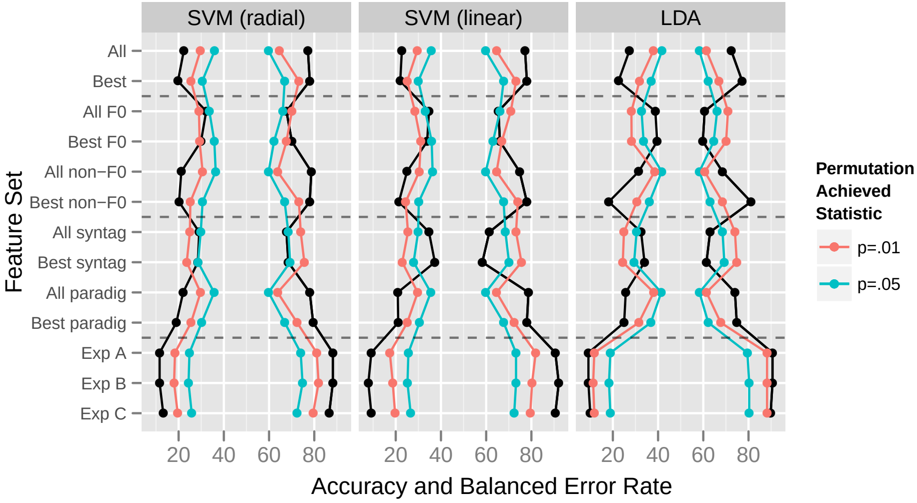

Figure 7: Accuracy and balanced error rates with permutation achieved significance for different classification models: training set lab; test set web2.

menter. The experimenter selected feature set included: duration of *I*, the first consonant closure duration of *did*, the F1-F2 differential at the midpoint of *I*, and the mean or maximum F_0 ratio between *I* and *did*. The performance of the classifiers which included a syntagmatic F_0 measure differed only marginally from the performance of the classifier which lacked either of the syntagmatic F_0 measures.

Thus, the performance of classifiers trained on laboratory data confirm that, while undeniably *relevant* acoustic cues, the F_0 and syntagmatic measures are not *necessary* acoustic cues for these data, contra prosodic-semantic theories of focus according to which F_0 (or pitch) and syntagmatic evaluation are unique or pre-eminent. In the following sections, we consider the contribution of individual measures in simpler, logistic regression models of production and perception.

With respect to contrasts between classifiers trained on laboratory data and classifiers trained on web data, we have not shown that there exists a statistically significant difference.[12] However, the consistent magnitude of difference, (i) between the web-trained lab-tested classifiers and the lab-tested web-trained classifiers and (ii) between web-trained web-tested classifiers and web-trained lab-tested classifiers (both around 5%), suggests an asymmetry. Both laboratory and web speech may be used as training data for web speech. However, the web speech proved somewhat less effective as training data for laboratory speech.

This asymmetry is consistent with the observation that many of the tokens from the web dataset were produced by professional broadcasters. These speakers are less likely to produce speech that is potentially ambiguous (e.g. produced with coarticulation, reduced vowels) and more likely to mark prosody consistently (Ostendorf & Shattuck-Hufnagel 1996) and with hyperarticulation. It follows then, that a classifier trained on clearer speech (with respect to these dimensions) would have more difficulty when applied to laboratory speech than on similarly clear speech; and a classifier trained on laboratory speech would do equally well on equally clear and less clear speech.

There is, of course, no *a priori* reason to expect this difference. Variation in the web speech–e.g. in terms of expressivity, recording quality, speaker, or discourse context–might have made it the superior training dataset instead of the laboratory data.

As a consequence, the web data may offer an important source of cross-validation, not only because they are produced more naturally and in a variety of different pragmatic contexts, but because the web speech appears to contain tokens with more idealized realizations. Moreover, it is encouraging that the classifiers work in either direction, as it suggests that results from lab speech lead to valid generalizations that extend to non-lab speech.

Finally, the failure of the base algorithm to identify the good features, and feature selection not doing as well as experimenter selection, is a negative result for a purely automatic procedure. This is a topic for further research using machine learning methodologies. Fresh data and more data will be a help in approaching it.

6 Model Comparison of Logistic Regression

6.1 Results

A direct statistical comparison of two SVM or LDA classifiers would require additional datasets, as discussed earlier. We may, however, get a sense of how individual features are contributing to less complex models, using logistic regression. In this section, we compare a logistic regression model of *web1* using experimenter-selected feature set A (‘duration_V2’, ‘duration_C3’, ‘f1f2Time50_V2’, ‘meanf0_ratio’) against four smaller models, each lacking a different member of the feature set, using an Analysis of Variance (ANOVA) (cf. Baayen et al. 2008). ANOVAs for the datasets *web1, web2* and *lab* in Table 5 show that the removal of any one of the four features is statistically significant. In other words, each of these features contributes significantly to explaining variation in the larger model.

6.2 Discussion

The classification experiments revealed the robustness of models with and without syntagmatic F_0 features. We observed informally that minimally contrastive feature sets–with and without a syntagmatic F_0 feature–did not yield large dif-

Table 5: Summary of Analysis of Variance (ANOVA) comparing a full logistic regression model with features ‘duration_V2’, ‘duration_C”, ‘f1f2Time50_V2’ and ‘f0_ratio’ against models lacking one of these features.

web1	Df	Residual Deviation	Df	Deviance	Pr (>Chisq)
‘duration_V2’, ‘duration_C3, ‘f1f2Time50_V2’, ‘f0_ratio’	85	19.266			
all except ‘duration_V2’	86	48.901	-1	-29.635	5.22E-08 *
all except ‘duration_C3’	86	25.917	-1	-6.6512	9.91E-03 *
all except ‘f2f2Time50_V2’	86	35.674	-1	-16.408	5.11E-05 *
all except ‘meanf0_ratio’	86	23.917	-1	-4.6515	0.03103 *

web2	Df	Residual Deviation	Df	Deviance	Pr (>Chisq)
‘duration_V2’, ‘duration_C3, ‘f1f2Time50_V2’, ‘f0_ratio’	122	44.301			
all except ‘duration_V2’	123	80.535	-1	-36.234	1.75E-09*
all except ‘duration_C3’	123	60.398	-1	-16.097	6.02E-05*
all except ‘f2f2Time50_V2’	123	53.216	-1	-8.9146	2.83E-03*
all except ‘meanf0_ratio’	123	53.498	-1	-9.1969	0.00242*

lab	Df	Residual Deviation	Df	Deviance	Pr (>Chisq)
‘duration_V2’, ‘duration_C3, ‘f1f2Time50_V2’, ‘f0_ratio’	389	208.51			
all except ‘duration_V2’	390	275.19	-1	-66.675	3.20E-16 *
all except ‘duration_C3’	390	221.46	-1	-12.949	3.20E-04 *
all except ‘f2f2Time50_V2’	390	250.28	-1	-41.772	1.03E-10 *
all except ‘meanf0_ratio’	390	216.36	-1	-7.8521	0.005076 *

ferences in classifier performance. The logistic regression models tested whether the ratio of F_0 means in *I* and *did* contributed meaningfully to a model with paradigmatic, non-F_0 measures.

Not only did the paradigmatic, non-F_0 features each contribute significantly to the logistic regression model, but the ratio of F_0 means contributed significantly as well. First, this result supports our conclusion from the classification experiments that both F_0 and non-F_0 and both syntagmatic and paradigmatic features are *relevant* to the classification of focus placement in these data. Second, this result is evidence against the hypothesis that the syntagmatic F_0 features are *redundant* (i.e. they contribute additional information) in these data. This result does not tell us, however, whether human listeners make use of syntagmatic, F_0 information in these data, a question we turn to in the final set of experiments.

7 Human acoustic classifiers

In this section, we assess the validity of the machine learning classifier results by comparing the machine learning classifiers to human classifiers. In other words, we want to know how closely the machine learning classifiers mimic human speech perception in classification accuracy and the acoustic measurements used to make judgements. We conducted two perceptual experiments to answer this question: the first using stimuli from the web dataset; the second using stimuli from the laboratory dataset.

7.1 Experiment 1: web stimuli

7.1.1 Method

A subset of 64 tokens from the *web2* corpus dataset was chosen: the first 32 of each semantic focus class. From each soundfile, the sequence "than I did" was extracted to create the stimulus. The files were normalized for sampling frequency and amplitude. The information presented to participants of the perception experiment was limited in this way in order to more closely match the limited information available to the statistical classifiers: neither machine nor human had the preceding or following acoustic information and neither machine nor human had any linguistic or extra-linguistic context.

Forty individuals participated in the perception experiment, which was conducted at the 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 stimulus, 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").

Of course, one may question whether a linguistically naïve participant can easily understand what 'prominent' means, and whether all participants in this experiment were indeed answering the same question. We note anecdotally, however, that participants seemed to find the task very natural and easy to complete, and given the results we have the impression that participants found the notion of prominence quite intuitive.

We evaluate the results in two ways. First, we calculate accuracy rates and balanced error rates, just as we did for the machine learning classifiers. In this way we can compare the human and machine learning classifiers using the same performance measures. We can also compare these measures by listener and by item. If many listeners consistently misclassified any of the data or any particular items were misclassified consistently, this would suggest a listener or item bias.

Second, we evaluate generalized linear mixed models using two of the top-performing feature sets. Mixed models allow us to incorporate random effects of listener and item. We chose the experimenter-selected feature sets *A* ('duration_V2', 'duration_C3', 'f1f2Time50_V2', 'mean_*f0*_ratio') and C ('duration_V2', 'duration_C3', 'f1f2Time50_V2'), because they were used for most of the top-performing machine learning classifiers, and because they differed in a single feature of interest, namely *mean_f0_ratio*. The modeling allows us to ask how much variance in *listeners' responses* a model using these features predicts. For a given model, we can also ask how predictive the individual features in the model are, and whether the model predicts significantly more variation than another model.

7.1.2 Results

Accuracy/Error

As a group, the 38 participants achieved a mean accuracy of 85.9%, median accuracy of 89.1% and standard deviation of accuracies 8.3%. They achieved a mean BER of 14.1%, median BER of 10.9% and standard deviation of BERs 8.3%. Participants' individual accuracy rates ranged from 64.1% to 95.3% and their balanced error rates ranged from 4.7% to 35.9% percent.

As for the items used in the experiment, only 3 were consistently misidentified by listeners. The majority of the stimuli were classified correctly more than 80% of the time. The mean by-item accuracy rate was 85.9%, the median 89.5% and the standard deviation 16.9%.

Generalized Linear Mixed Models

In order to understand which acoustic features listeners were using to make their judgments, we tested for the statistical significance of individual features in generalized linear mixed models using the R package lme4 (Bates et al. (2007)).

Both statistical models were significant, as were each of the individual fixed effects (i.e. the acoustic features), with the notable exception of *mean_f0_ratio* (cf. Table 6).

We can quantify whether one of the two models of listener response is more predictive than the other using ANOVA. The various goodness of fit criteria (AIC, BIC and log likelihood) for our two models are very similar and according to the $\chi 2$ test statistic, we cannot conclude that the model using feature set

Table 6: Summary of generalized linear mixed models for listener responses to a subset of *web2* using predictors from hand-selected feature sets *Experimenter-selected A* and *Experimenter-selected C*. Test statistic Wald z-score; statistical significance (p<0.01) indicated by asterisks.

Generalized Linear Mixed Model of Listener Response (Web Data)

EXPERIMENTER-SELECTED A: duration_V2, duration_C3, f1f2Time50_V2, mean_*f0*_ratio

Random effects:

Groups	Variance	Std. Dev.
Participant	0.066720	0.25830
Item	0.041699	0.20420

Fixed effects:

	Estimate	Std. Error	z-value	p-value
Intercept	1.249	0.5746	2.174	0.0297 *
duration of *I*	36.05	2.332	15.457	<2e-16 *
duration of first closure in *did*	-45.24	3.509	-12.893	<2e-16 *
F1F2 differential at midpoint of *I*	-0.003067	0.0003291	-9.318	<2e-16 *
ratio of mean F_0 in *I* and *did*	0.004232	0.02654	-0.159	0.873 **n.s.**

EXPERIMENTER-SELECTED C: duration_V2, duration_C3, f1f2Time50_V2

Random effects:

Groups	Variance	Std. Dev.
Participant	0.066720	0.25830
Item	0.041699	0.20420

Fixed effects:

	Estimate	Std. Error	z-value	p-value
Intercept	1.210236	0.520745	2.324	0.0201 *
duration of *I*	35.946678	2.254716	15.943	<2e-16 *
duration of first closure in *did*	-45.078265	3.401762	-13.251	<2e-16 *
F1F2 differential at midpoint of *I*	-0.003068	0.000329	-9.326	<2e-16 *

Experimenter-selected A predicts significantly more variation than the model using feature set *Experimenter-selected C*. In other words, we cannot say that adding the feature *mean_f0_ratio* results in a more predictive model of listeners' responses.

We can also perform model comparison to assess the contribution of the random effects: participant and item. A model with participant and item as random effect and *Experimenter-selected A* as fixed effect explains significantly more variation than a model with participant only as random effect and featureset *Experimenter-selected A* as fixed effect ($\chi 2$=80.533, p=2.2e-16). Similarly, a model with participant and item as random effect and *Experimenter-selected A* as fixed effect explains significantly more variation than a model with item only as random effect and featureset *Experimenter-selected A* as fixed effect ($\chi 2$=21.465, p=3.603e-16).

Confidence Rating

Participants' confidence rating turned out to be a very significant predictor of their performance on a given stimuli (generalized linear model: σ= 0.031, z= -10.81,p<0.001). This indicates that listeners have a degree of introspective access to gradience or ambiguity in the prominence distinction.

7.1.3 Discussion

The performance of listeners in the perception experiment, as measured by classification accuracy and balanced error rate, closely matched that of the machine learning classifiers. Recall that the top-performing classifier achieved an accuracy rate of 92.9%: while some listeners' accuracy rates were as low as 64%, 16 out of the 38 human classifiers achieved an accuracy rate above 90%.[13]

The comparison of listener response models revealed that item explains a statistically significant amount of listener variation. Review of the item distribution, however, reveals that 3 of the 64 items were effectively outliers, with accuracy rates well below 50%. The poor human classifier performance on these items suggests that misclassification by the machine learning classifiers are likely to be a result of other variation (e.g. speaker disfluency, high signal-to-noise ratio) in the data by which human listeners were equally misled.

One misclassified example, transcribed in (22), received a listener accuracy rate of 18.4%. The co-reference criterion predicts this example will be realized with subject focus since the subjects of the two clauses do not co-refer; however, the matrix clause also has a salient contrast 'at that time' which licenses focus on *did.*

(22) Growing up at that time and that location, you can't have more fun as a kid

than [I]$_{\mathbf{F}}$ [did]$_{\mathbf{F}}$

Example (22) is an infrequently occurring but linguistically possible example of double focus. The task of the machine learning classifiers and the human listeners was binary (two semantic classes for the machine learning classifiers and two prominence choices for the human listeners), while example (22) effectively belongs to a third class.

Finally, the same feature sets used in the top-performing machine learning classifiers (viz. *Experimenter-selected A* and *Experimenter-selected C*) were statistically significant in a model of listener response. There was no main effect for the mean F_0 ratio feature (i.e. it was not individually significant in the model), and removing the feature did not result in a less explanatory model. This result is consistent with the corresponding machine learning classifiers, for which the addition of the feature *mean_f0_ratio* did not substantially improve generalization accuracy or error rates.

7.2 Experiment 2: laboratory production stimuli

7.2.1 Methodology

In the second perception experiment, human listeners were presented with excerpts of "than I did" taken from a subset of the laboratory production data.[14] The experiment was carried out with the same methodology as in Experiment 1. Forty-one individuals participated.

7.2.2 Results

Accuracy/Error

The human acoustic classifiers performed on par with the machine learning classifiers. For the lab data, the 41 listeners achieved a mean accuracy of 78.5%, median accuracy 81.3% and standard deviation of accuracies 13.1%. They achieved a mean balanced error rate of 13.1%, median balanced error rate of 12.2% and a standard deviation of balanced error rates 6.9%. Participants' individual accuracy rates ranged from 53.1% to 96.9% and their balanced error rates ranged from 3.7% to 29.3%.

As for the items used in the experiment, 1 lab item was correctly identified at less than 50%. Among lab items, the mean accuracy rate was 78.5%, median 80.5% and the standard deviation 16.9%.

Stimuli were drawn from eight different speakers in the production experiment. The accuracy rates for individual speakers ranged from 67.4 to 82.0%. The mean accuracy rate among speakers was 74.8%, median 73.9% and standard deviation 5.0%.

Generalized Linear Mixed Models

As in the first perception experiment, in order to understand which acoustic features listeners were using to make their judgments, we evaluated generalized linear mixed models using two top-performing feature sets. *Experimenter-selected A* contains the features 'duration_V2', 'duration_C3' and 'f1f2Time50_V2' as fixed effects; *Experimenter-selected C* contains the features 'duration_V2', 'duration_C3', 'f1f2Time50_V2' and 'mean_f0_ratio' as fixed effects. The two models differed only in the feature 'mean_f0_ratio'. Both models also contained participant and item as random effects.

All of the listener response models were statistically significant. There were main effects for each of the acoustic features, with the notable exception of 'mean_f0_ratio', which was not significant. The feature 'duration_V3' was only marginally significant in the lab model using feature set *Experimenter-selected A*.

Despite a marginal test statistic for the 'f0_ratio' parameter estimate in the lab model (p=0.23795 is small but above an acceptable rate of α=0.05), an ANOVA comparing the two lab models with and without 'mean_f0_ratio' suggests that the addition of this feature does indeed result in a more predictive model of listener response (χ2=80.533, p=2.2e-16).

ANOVAs revealed that the addition of item as random effect resulted in a more predictive model of listener response on the *lab* dataset (χ2=13.654, p=0.0002198). There was not sufficient evidence to conclude that including participant as random effect resulted in more predictive model.

Confidence Rating

Participants' confidence rating turned out to be a very significant predictor of their performance on a given stimulus (generalized mixed-effects linear model: σ= 0.05844, z= 7.429, p< 1.10e-13), indicating introspective sensitivity to the reliability of classification judgments.

7.2.3 Discussion

The performance of listeners in the perception experiment, as measured by classification accuracy and balanced error rate, was on par with that of the machine learning classifiers.

Both logistic regression models confirmed that the contribution of the paradigmatic, non-F_0 measures was statistically significant. There was insufficient evidence that the feature 'mean_f0_ratio' contributed significantly (p=0.24). However, an ANOVA comparing the two models–one with and one without 'mean_f0_ratio'–revealed that the feature does in fact explain a statistically significant amount of variation in listener response which the other variables in the model do not. This suggests that listeners are using both the syntagmatic, F_0 and paradigmatic non-F_0 measures in the lab data.

As in the perception experiment using web-harvested data, listeners did have difficulty with a handful of items. The ANOVA comparing models with and without item as a random effect was significant, indicating that item explains a statistically significant amount of variation. The listeners' less than perfect performance, like the machine classifiers' performance, may be at least partly explained by these outliers.

8 Conclusion

8.1 Discussion of results

We set out to test predictions of theories of focus interpretation in one constrained environment. According to an anaphoric theory of focus, the location of focus in the comparative clause is determined by the matrix clause. Operatively, the location of prominence can be predicted according to the (co-)reference of the subjects in the main and comparative clauses (cf. 4). The machine learning experiments confirmed the robustness of this generalization with both naturally occurring and experimentally elicited data. Classifiers trained exclusively on acoustic measurements from web-harvested data achieved accuracy rates as high as 92.9% and balanced error rates as low as 6.5% when tested on similar web-harvested data. They achieved accuracy rates as high as 87.6% and error rates as low as 10.5% when tested on laboratory-elicited data, still well above a baseline of 51.0% accuracy. Classifiers trained exclusively on acoustic measurements from laboratory-elicited data achieved accuracy rates as high as 89.0% and balanced error rates as low as 10.5% when tested on web-harvested data.

The human classification experiments confirmed the robustness of the generalization as well. Listeners presented only with web-harvested tokens of 'than I did' achieved a mean classification accuracy of 86.4% (standard deviation 8.1%) and a mean balanced error rate of 4.5% (standard deviation 2.8%). Listeners presented only with laboratory-elicited tokens of 'than I did' achieved a mean accuracy rate of 78.5% (standard deviation 13.1%) and a mean balanced error rate of 13.1% (standard deviation 6.8%).

In building the classifiers, we also took the opportunity to compare the contribution of specific groupings of phonetic measures: one division between F_0 and non-F_0measures and one division between paradigmatic and syntactic measures. In the machine learning classification, we observed a tendency for classifiers with exclusively non-F_0 features to meet or exceed the performance of exclusively F_0 features. Of course, we do not wish to suggest that F_0 never signals prosodic prominence, as this is well attested. Indeed, classifiers using exclusively F_0 measures achieved accuracies as high as 79.5% and it is certainly possible that that improved acoustic modeling of F_0 may yield even better classifier performance. We leave detailed examination of the F_0 profile for future study but note informally that while many tokens with expected focus on *I* appeared to have H* pitch accents, we also observed the occurrence of other intonational patterns in which the F_0 maximum of *I* was exceeded by the F_0 maximum of *did* (e.g. Figures 8, 9, 10) or in which the F_0 maximum was delayed (e.g. Figure 11). Human listeners may well be able to recover focus despite such variability.

Rather than interpreting the results as indicating non-relevance of F_0, we wish to highlight the contribution of non-F_0 measures, which turned out to be highly predictive. Non-F_0 measures, we argue, offer practical benefits for automatic detection of focus. We also note that many researchers have taken such findings not merely as evidence for the existence of secondary cues of accent, but as evidence against the pre-eminence of pitch accent (e.g. Kochanski 2006; Fant et al. 1991; Sluijter & van Heuven 1996; Heldner et al. 1999; Heldner 2003; among others). Mo (2010) finds that individuals show considerable variation in which combinations of acoustic measures they use to mark prominence and these combinations include F_0 to varying degrees.

We propose that the increased duration and especially vowel quality ob-

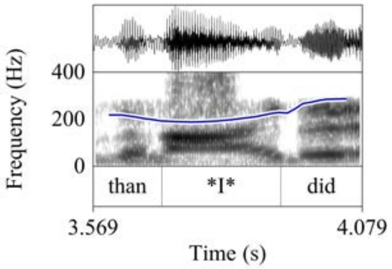

Figure 8: Waveform, spectrogram, and F_0 contour from excerpt of web file 117

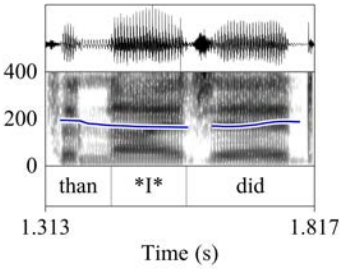

Figure 9: Waveform, spectrogram, and F_0 contour from excerpt of lab recording 327_6

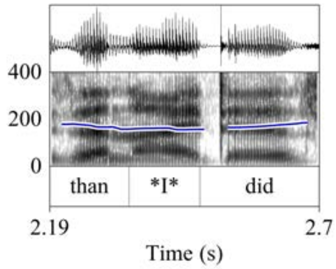

Figure 10: Waveform, spectrogram, and F_0 contour from excerpt of web file 322_10

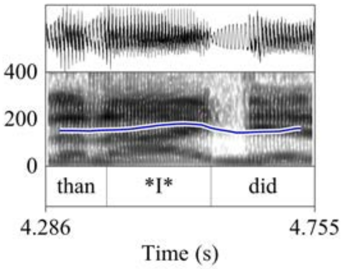

Figure 11: Waveform, spectrogram, and F_0 contour from excerpt of lab recording 327_6

served in our data correlate with post-lexical or utterance-level stress. Following the autosegmental-metrical tradition, stress is phonologically distinct from pitch accent (e.g. Liberman 1975; Pierrehumbert 1980), although related by the requirement for pitch accent to align with utterance-level stress.

For the second grouping of measures, we observed a tendency for classifiers with exclusively paradigmatic measures to meet or exceed the performance of exclusively syntagmatic measures. Again, we emphasize that syntagmatic measures are indeed predictive; classifiers using exclusively syntagmatic measures achieved accuracies as high as 80.3% and it is entirely possible that inclusion of other syntagmatic measures might have yielded even better classifier performance. Rather, we highlight the contribution of the paradigmatic measures, which turned out to be highly predictive.

As described in the Introduction, it is traditionally held that prosodic prominence is relational or "syntagmatic", meaning that prominence is processed relative to the sentence that is being uttered (e.g. Jakobson et al. 1952; Trubetzkoy 1939; Lehiste 1970; Ladefoged 1975; Hyman 1978). This explains, among other phenomena, how a word may be perceived as prominent in either fast or slow speech.

Segmental phenomena such as vowel quality, voice quality and, in some cases, duration are "paradigmatic", meaning that they are processed relative to another possible realization. Segmental phenomena, such as the phonological voicing contrast between [p] and [b] are responsible for meaning-distinguishing minimal pairs like *pig* and *big*. There are no minimal pairs in English, so the reasoning goes, that are distinguished solely by pitch (e.g. *pig* with a high tone and *pig* with a low tone).

Minimal prosodic pairs (or n-tuples) do exist, however, as we've seen (cf. 3 a,3 b,3 c). How can we understand these essentially paradigmatic contrasts without also denying the syntagmatic character of prosodic prominence? Within metrical stress theory, prosody is hierarchical, and one can speak of prominence at multiple levels. Prominence at the word level is realized phonologically by stress, and it is possible to distinguish individual words using stress (e.g. *ímport* vs. *impórt*). Further, one can make intonational contrasts at the phrase or utterance level.

Phonologically, then, the difference between two minimal intonational pairs is thus both syntagmatic—how prosodic elements are grouped and which prosodic element is most prominent within a grouping—and paradigmatic—how the prosodic structure of one utterance differs minimally from the prosodic structure of another.

An important source of evidence against uniquely syntagmatic accounts comes from cases of double focus. Ladd (1991) describes an individual "who used to be able to speak German well but then had then spent a long time living in Sweden and now spoke good Swedish but had trouble with German". Ladd replies to the individual with (23).

(23) **That's what happened to MY FRENCH.**

It used to be good, but then I spent a year in Germany and ended up with good German, and now whenever I want to speak French I get German interference all over the place.

Semantically, (23) is a case of double focus, on *my* and on *French.* And phonologically, this focus is being conveyed with prominence. Ladd observes, however, that prominence on *my* cannot be purely syntagmatic. It is not the case that my is more prominent than its sister, *French*; if anything, *French* is realized with greater prominence than is *my.* The necessary comparison is paradigmatic: (23) is compared to the minimally different realization in (24).

(24) **That's what happened to my FRENCH.**

Similarly, measures of prominence on *I* alone were good unique predictors in the *than I did* datasets because the salient contrast was not only syntagmatic, but paradigmatic: i.e. between focal and non-focal realizations of *I.*

(25) than [I]$_{(F)}$ did   *paradigmatic contrast*
↕
than I (did)$_F$ (. . .) $_F$

As we've noted, the highly predictive paradigmatic measures in our data were also measures of stress, namely duration and vowel quality. We hypothesize that this is largely due to the lexico-syntactic class of the focused constituent: i.e. function words tend to be unaccented unless focused. Ladd's examples (23,26) contrast focal and non-focal realizations of the function word *my.* In our comparatives data, we are contrasting focal and non-focal realizations of the function word *I.* It is sufficient for the usually non-prominent pronoun to indicate prominence by realizing it with even a low degree of prominence.

It is well known that there are important phonological distinctions between function words and lexical words (e.g. Selkirk 1996and references cited therein) and lexical words may require a greater degree of prominence to signal semantic focus. Ladd offers another prosodic minimal pair in which the prosodic contrast is realized on the lexical word *butcher.* In this well-known example, *butcher* is understood as an epithet for surgeon when unfocused, and literally as a butcher when focused.

(26) a. A: Everything OK after your operation?

B: Don't talk to me about it.

The **butcher** charged me a thousand BUCKS!   *epithet*

b. A: Everything OK after your operation?

B: The **BUTCHER** charged me a thousand bucks!   *literal*

Ladd intuits that the prosodic contrast in (23-24) is not equivalent to the contrast in (26a-26b). For the pronouns, vowel reduction appears to be sufficient to mark the distinction, while both of the lexical words have unreduced vowels and a contrast in pitch appears to be necessary.

In the *thanIdid* datasets, the robustness of measures which are non-intonational and which are extracted only from *I* reflects the categorical and largely paradigmatic prominence on focused *I*. A full, unreduced vowel, as indicated phonetically by greater duration and more extreme formant extrema, is sufficient information to identify the function word as focused with considerably accuracy. It is likely the case that humans use a combination of syntagmatic and paradigmatic information, and that the choice is context-dependent. Mo (2010), much like the Boruta feature selection in this study, finds that listeners tend to make use of paradigmatic duration and formant information, but syntagmatic loudness measures.

The category of focus as it figures in current theory can be characterized as a grammatically mediated correlation between a semantics-pragmatics of contrast and redundancy, and a phonetics of prominence. The positive results obtained here suggest the feasibility of constructing explicit numerical models of this correlation using machine learning, and of testing the predictions of formalized theories of information structure in data collected in the "wild" of spoken language used on the web.

The web methodology that is detailed in Howell & Rooth (2009) retrieves tokens of specific lexical strings. In related work, we collected data for several dozen word-string targets and filtered and transcribed the results (Rooth et al. 2013; Lutz et al. 2013). Structuring datasets around specific target strings is a limitation, but it is also an advantage in that it allows machine learning to use specific features in the target. Investigating the success of the method for other contexts, and generalizing the method to an open-ended class of contexts for focus realization is a topic for future research.

Acknowledgments

We wish to thank audiences at McGill, Cornell, Montclair State, York and Concordia universities, the 1st Digging into Data conference, Experimental and Theoretical Advances in Prosody (ETAP) conference, the Association for Canadian Studies in the United States (ACSUS), the Montréal-Ottawa-Toronto Phonology Workshop, New Tools and Methods for Very-Large-Scale Phonetics Research and the Acoustic Society of America. We also thank the RAs in the prosodylab at McGill for their help in running the experiment and processing the results. We are also grateful to Associate Editor Jonathan Barnes and two anonymous reviewers. This research was supported by the following grants: NSF 1035151 RAPID: Harvesting Speech Datasets for Linguistic Research on the Web (Digging into Data Challenge), SSHRC Digging into Data Challenge Grant 869-2009-0004, the SSHRC CRC program and SSHRC 410-2011-1062 Relative prosodic boundary strength and its role in encoding syntactic structure.

Notes

[1] The authors do, of course, also allow for cases of focus within constituents smaller than a proposition. The reader is referred to these works for the full proposals.

[2] Here the degree variable is existentially quantified. The same results are obtained if there are occurrences of the same free degree variable in the main clause and the comparative clause.

[3] Although the co-reference criterion divides instances of the comparative exhaustively, it should be noted that there are certain cases in which it does not correspond exactly with theoretical accounts of focus. In particular, the co-reference criterion does not distinguish cases of double focus, such as (i). The co-reference criterion predicts that (i) belongs in class "s" (subject focus).

(i) You should have earned less last year than $[\mathrm{I}]_{\mathrm{F}}$ did $[\mathrm{this}]_{\mathrm{F}}$ year Antecedent: You should have earned x much last year 'You should have earned x much last year' entails 'someone earned x much at some time'.

[4] The *Everyzing* and *play.it* interfaces are no longer available, although the same technology has since been made available for a variety of different content providers, including WNYC, Fox Business and PBS. For tools which interact with these newest interfaces, readers are invited to contact the authors.

Researchers interested in incorporating web-harvested speech corpora are also advised to identify and review other similar resources appropriate for their needs. At the time of writing, we are aware of at least three additional sources of transcribed and time-indexed naturally-occurring speech available on the web.

- Audiosear.ch and its associated API provides full-text search of podcasts including programming from National Public Radio (NPR) and the Canadian Broadcasting Corporation (CBC). Many ASR-generated transcripts are manually corrected and so transcription quality can be quite excellent.
- Google and YouTube host videos with closed captioning/subtitles which are time-aligned to short stretches of speech.
- Digital artist Sam Levine has provided python scripts Levine (2016) which search subtitle files from YouTube and other media or from transcriptions generated by the open-source automatic speech recognition system CMU Sphinx (Lamere et al. (2003); Huggins-Daines et al. (2006)).

Levine's platform allows for GREP search of regular expressions and part of speech tagging, which could be applied, for example, to focus-sensitive expressions which are themselves discontinuous (e.g. either... or) from their focus associate (e.g. only).

[5] As noted in Footnote 4, the audio search landscape has evolved since we first collected our web data and we were limited to 90 and 127 tokens in our training and test sets, respectively.

[6] The sample size should be 10 times the number of attributes according to Brown & Tinsley (1983), 20 times the number of attributes according to Stevens (2002).

[7] The methods of regularized discriminant analysis (Friedman (1989)) or shrinkage discriminant analysis (Ahdesmäki & Strimmer (2010)) have been proposed to improve performance of simple discriminant analysis when the number of attributes exceeds the size of the dataset. We do not pursue these methods here.

[8] The two and three-dimensional models underlying Figures 1 and 2 were obtained the MATLAB *fitcdiscr* function.

[9] The terms 'feature' and 'attribute' are used here in their statistical or computational sense, referring to a particular vector of data (e.g. the vector of data corresponding to 2nd vowel duration). Note also that the terms 'feature' and 'attribute' are often used to distinguish predictors before and after kernel mapping, respectively. Since nothing in the study hangs on this distinction, we will use the terms interchangeably.

[10] This is the typical case; however, as an anonymous reviewer notes, it is not a guaranteed effect.

[11] Equation for RBF kernel: $K(x, x^0) = exp(-\gamma \left\| x - x^0 \right\|^2$

[12] McNemar's test does not measure variability due to training set.

[13] The task of the human and of the machine were similar in that both had access to only the acoustic information from the string "than I did". Although we are encouraged by the close

match in performance, we must also note that it is possible that a human may achieve greater performance on a task more closely related to how they usually use language, as opposed to the metalinguistic task used here of identifying prominence. We leave this for future research.

[14] We used speech from the first 8 participants of the production study. We used 8 of the original 16 elicited utterances—the same 8 for each of the 8 speakers: tokens 1, 3, 5, 7, 9, 11, 13 and 15.

References

Ahdesmäki, Miika & Korbinian Strimmer. 2010. Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *The Annals of Applied Statistics* 4(1). 503–519. doi:10.1214/09-AOAS277.

Baayen, RH, DJ Davidson & DM Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412. doi:10.1016/j.jml.2007.12.005.

Barron, Andrew R. 1994. Approximation and estimation bounds for artificial neural networks. *Machine Learning* 14(1). 115–133.

Bartels, Christine. 2004. Acoustic correlates of 'second occurrence' focus: Towards an experimental investigation. In Hans Kamp & Barbara H. Partee (eds.), *Context-dependence in the analysis of linguistic meaning*, 354–361. Amsterdam: Elsevier.

Bates, Douglas, Deepayan Sarkar, Maintainer Douglas Bates & LinkingTo Matrix. 2007. The lme4 package. *R package version* 2(1).

Beaver, David, Brady Zack Clark, Edward Flemming, T Florian Jaeger & Maria Wolters. 2007. When semantics meets phonetics: Acoustical studies of second-occurrence focus. *Language* 83. 245–276. doi:10.1353/lan.2007.0053.

Beaver, D.I. & B.Z. Clark. 2008. *Sense and sensitivity: How focus determines meaning*. Wiley-Blackwell. doi:10.1002/9781444304176.

Beckman, Mary E & Gayle M Ayers. 1994. Guidelines for tobi labeling guide, version 2.0.

Beckman, Mary E & Jan Edwards. 1994. Articulatory evidence for differentiating stress categories. In P. Keating (ed.), *Papers in laboratory phonology iii: Phonological structure and phonetic form*, vol. 3, 7–33. Cambridge: Cambridge University Press.

Beckman, Mary E. & Janet B. Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3. 255–310.

Bishop, Jason B. 2008. The effect of position on the realization of second occurrence focus. In *Proceedings of interspeech*, International Speech Communications Association.

Boersma, Paul & David Weenink. 2013. PRAAT, a system for doing phonetics by computer. Computer Program.

Bolinger, D. 1958. Stress and information. *American Speech* 33. 5–20. doi: 10.2307/453459.

Bolinger, Dwight Le Merton. 1981. *Two kinds of vowels, two kinds of rhythm.* Reproduced by the Indiana University Linguistics Club.

Boser, Bernhard E, Isabelle M Guyon & Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory*, 144–152. ACM. doi: 10.1145/130385.130401.

Breen, Mara, Evelina Fedorenko, Michael Wagner & Edward Gibson. 2010. Acoustic correlates of information structure. *Language and Cognitive Processes* 25(7). 1044–1098. doi:10.1080/01690965.2010.504378.

Brown, Michael T & Howard E Tinsley. 1983. Discriminant analysis. *Journal of Leisure Research* 15. 290–310.

Brown, Michael T & Lori R Wicker. 2000. *Discriminant analysis.* San Diego: Academic Press. doi:10.1016/b978-012691360-6/50009-4.

Bruce, Gösta. 1977. *Swedish word accents in sentence perspective.* Lund: Gleerup.

Campbell, N. & M. Beckman. 1997. Stress, prominence, and spectral tilt. In Antonis Botinis, Georgios Kouroupetroglou & George Carayiannis (eds.), *Intonation: Theory, models and applications (proceedings of an esca workshop, september 18-20, 1997, athens, greece).*, 67–70. ESCA and University of Athens Department of Informatics.

Chang, Chih-Chung & Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3). 27. doi:10.1145/1961189.1961199.

Cho, Taehong. 2006. Manifestation of prosodic structure in articulation: Evidence fom lip kinematics in English. In L. M. Goldstein, D. H. Whalen & C. T. Best (eds.), *Laboratory phonology 8: Varieties of phonological competence*, 519–548. Berlin/New York: Mouton de Gruyter.

Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English.* New York: Harper & Row.

Cortes, Corinna & Vladimir Vapnik. 1995. Support vector networks. *Machine learning* 20(3). 273–297. doi:10.1007/BF00994018.

Cristianini, Nello & John Shawe-Taylor. 2000. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge University Press. doi:10.1017/cbo9780511801389.

De Jong, Kenneth J. 1991. *The oral articulation of english stress accent*: Ohio State University dissertation.

Dietterich, Thomas G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10(7). 1895–1923.

Dimitriadou, Evgenia, Kurt Hornik, Friedrich Leisch, David Meyer & Andreas Weingessel. 2009. e1071: Misc functions of the department of statistics (e1071), tu wien v1. 5-19. TU Wien.

Drubig, Hans Bernhard. 1994. Island constraints and the syntactic nature of focus and association with focus. In *Arbeitspapiere des Sonderforschungsbereichs 340: Sprachtheoretische Grundlagen der Computerlinguistik*, vol. 51, Tübingen/Stuttgart: Sonderforschungsbereich 430.

Drubig, Hans Bernhard. 2003. Toward a typology of focus and focus constructions. *Linguistics* 41. 1–50. doi:10.1515/ling.2003.003.

Evgeniou, T., M. Pontil & T. Poggio. 2000. Regularization networks and support vector machines. *Advances in Computational Mathematics* 13. 1–50. doi: 10.1023/A:1018946025316.

Fant, Gunnar, Anita Kruckenberg & Lennart Nord. 1991. Durational correlates of stress in Swedish, French and English. *Journal of Phonetics* 19. 351–365.

Friedman, Jerome H. 1989. Regularized discriminant analysis. *Journal of the American Statistical Association* 84. 165–175.

Fry, D.B. 1955. Duration and Intensity as Physical Correlates of Linguistic Stress. *Journal of the Acoustical Society of America* 27. 765–768.

Fry, D.B. 1958. Experiments in the perception of stress. *Language and Speech* 1(2). 126–152.

Giegerich, Heinz. 1985. *Metrical phonology and phonological structure: German and english*. Cambridge University Press.

Gillick, Laurence & Stephen J Cox. 1989. Some statistical issues in the comparison of speech recognition algorithms. In *Acoustics, speech, and signal processing, 1989. icassp-89., 1989 international conference on*, 532–535. IEEE.

Goldsmith, John A. 1976. *Autosegmental phonology*: MIT dissertation.

Gorman, Kyle, Jonathan Howell & Michael Wagner. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics* 39(3). 192–193.

Grefenstette, Gregory. 1999. The world wide web as a resource for example-based machine translation tasks. *Translating and the Computer* 21. 4–7.

Gussenhoven, Carlos. 1992. Sentence accents and argument structure. In I. Roca (ed.), *Thematic structure: Its role in grammar*, 79–106. Foris.

Gussenhoven, Carlos. 2004. *The phonology of tone and intonation*. Cambridge: Cambridge University Press. doi:10.1017/cbo9780511616983.

Halle, Morris & Jean-Roger Vergnaud. 1987. *An essay on stress*. Cambridge: MIT Press.

Halliday, M.A.K. 1967. *Intonation and grammar in british english*. Mouton The Hague. doi:10.1515/9783111357447.

Hayes, Bruce. 1981. *A metrical theory of stress rules*: MIT dissertation.

Heldner, Mattias. 2003. On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics* 31. 39–62. doi:10.1016/S0095-4470(02)00071-2.

Heldner, Mattias, Eva Strangert & Thierry Deschamps. 1999. A focus detector using overall intensity and high frequency emphasis. In *Proceedings of the international congress of phonetic sciences*, vol. 99, 1491–1493. San Francisco: Linguistics Department, University of California Berkeley.

Howell, Jonathan. 2011. Second occurence focus and the acoustics of prominence. In C. Ulbrich R. Folli (ed.), *Interfaces in linguistics: New research perspectives*, 278–298. Oxford University Press.

Howell, Jonathan. 2016. Replication data for: Acoustic classification of focus: on the web and in the lab. Harvard Dataverse, V1. doi:10.7910/DVN/BGCNPE.

Howell, Jonathan & Mats Rooth. 2009. Web harvest of minimal intonational pairs. In S. Sharoff I. Alegria, I. Leturia (ed.), *Proceedings of the fifth web as corpus workshop*, 45–52. San Sebastian, Spain: Elhuyar Fundazioa.

Hsing, Tailen, Sanju Attoor & Edward Dougherty. 2003. Relation between permutation-test p values and classifier error estimates. *Machine Learning* 52. 11–30. doi:10.1023/A:1023985022691.

Hsu, Chih-Wei, Chih-Chung Chang & Chih-Jen Lin. 2003. A practical guide to support vector classification. National Taiwan University. https://www.cs.sfu.ca/people/Faculty/teaching/726/spring11/svmguide.pdf.

Huberty, Carl J. 2006. *Applied manova and discriminant analysis*. Hoboken, N.J.: Wiley-Interscience 2nd edn. doi:10.1002/047178947x.

Huggins-Daines, David, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar & Alexander I Rudnicky. 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *International Conference on Acoustics Speech and Signal Processing Proceedings*, 1–185.

Hyman, L.M. 1978. Tone and/or accent. *Elements of tone, stress and intonation* 1. 20.

Jacobs, Joachim. 1999. Informational autonomy. In P. Bosch & R. van der Sandt (eds.), *Focus: Linguistic, cognitive, and computational perspectives*, 56–81. Cambridge University Press.

Jakobson, R., C.G.M. Fant & M. Halle. 1951. Preliminaries to speech analysis (cambridge, mass.). *First printing* .

Jakobson, Roman, Gunnar Fant & Morris Halle. 1952. *Preliminaries to speech analysis*. Cambridge, MA: MIT Press.

Jensen, David. 1992. *Induction with randomization testing: decision-oriented analysis of large data sets*. St. Louis, Missouri: Washington University dissertation.

Joachims, Thorsten. 1997. A probabalistic analysis of the rocchio algorithm with tfidf for text categorization. In *Proceedings of international conference on machine learning (icml)*, .

Kadmon, Nirit. 2001. *Formal pragmatics: Semantics, pragmatics, presupposition, and focus*. Malden, MA: Blackwell.

Keerthi, S Sathiya & Chih-Jen Lin. 2003. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural computation* 15(7). 1667–1689. doi:10.1162/089976603321891855.

Kilgarriff, Adam. 2007. Googleology is bad science. *Computational Linguistics* 33(1). 147–151. doi:10.1162/coli.2007.33.1.147.

Klecka, William R. 1980. *Discriminant analysis*. Beverly Hills: Sage Publications.

Kochanski, Greg. 2006. Prosody beyond fundamental frequency. In Stefan Sudhoff, Denisa Lenertova, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter & Johannes Schließer (eds.), *Methods in empirical prosody research*, 89–122. Berlin; New York: Walter de Gruyter. doi: 10.1515/9783110914641.89.

Kursa, Miron B. & Witold R. Rudnicki. 2010. Feature selection with the boruta package. *Journal of Statistical Software* 36. 1–13. doi:10.18637/jss.v036.i11.

Lachenbruch, Peter. 1975. *Discriminant analysis*. New York: Hafner 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*, vol. 7, .

Ladefoged, Peter. 1967. *Three areas of experimental phonetics*. London: Oxford University Press.

Ladefoged, Peter. 1975. *A course in phonetics*. Orlando: Harcourt Brace.

Ladefoged, Peter & Gerald Loeb. 2002. Preliminary studies on respiratory activity in speech. *UCLA Working Papers in Phonetics* 101(50-60).

Lamere, Paul, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth & Peter Wolf. 2003. The CMU SPHINX-4 speech recognition system. In *International conference on acoustics, speech and signal processing, Hong Kong*, 2–5.

Leben, William Ronald. 1973. *Suprasegmental phonology.*: Massachusetts Institute of Technology dissertation.

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

Levine, Sam. 2016. Github repository. *VideoGrep* http://antiboredom.github.io/videogrep.

Liberman, Mark Y. 1975. *The intonational system of English*: MIT dissertation.

Liberman, Mark Y. & Alan S. Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8. 249–336.

Lin, H.T. & 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.

Lutz, David, Parry Cadwallader & Mats Rooth. 2013. A web application for filtering and annotating web speech data. In *Web as Corpus 8*, ACL SIGWAC. Code at https://github.com/del82/ezra.

Lyons-Weiler, J., R. Pelikan, H.J. Zeh III, D.C. Whitcomb, D.E. Malehorn, W.L. Bigbee & M. Hauskrecht. 2005. Assessing the statistical significance of the achieved classification error of classifiers constructed using serum peptide profiles, and a prescription for random sampling repeated studies for massive high-throughput genomic and proteomic studies. *Cancer Informatics* 1(1). 53.

McNemar, Quinn. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2). 153–157.

Mo, Yoonsook. 2010. *Prosody production and perception with conversational speech*: University of Illinois at Urbana-Champaign dissertation.

Molinaro, Annette M, Richard Simon & Ruth M Pfeiffer. 2005. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 21(3301-3307). doi:10.1093/bioinformatics/bti499.

Mukherjee, Sayan, P Tamayo, D Slonim, A Verri, T Golub, J Mesirov & T Poggio. 1999. Support vector machine classification of microarray data. *AI Memo 1677, Massachusetts Institute of Technology* .

Naes, Bjorn-Helge Mevik, Tomod. 2001. Understanding the collinearity problem in regresssion and discriminant analysis. *Journal of Chemometrics* 15(413-426). doi:10.1002/cem.676.

Ostendorf, Patti Price, Mari & Stefanie Shattuck-Hufnagel. 1996. *Boston university radio speech corpus*. Philadelphia.

Partee, H., Barbara. 1991. Topic, focus and quantification. In *Proceedings of SALT I*, Cornell, Ithaca, NY: CLC Publications. doi:10.3765/salt.v1i0.2918.

Pierrehumbert, Janet. 1980. *The phonology and phonetics of english intonation*: MIT dissertation. doi:http://hdl.handle.net/1721.1/16065.

Prince, Alan S. 1983. Relating to the grid. *Linguistic Inquiry* 14. 19–100.

R Development Core Team. 2013. *R: A language and environment for statistical computing*. http://www.R-project.org.

Rooth, Mats. 1992. A theory of focus interpretation. *Natural Language Semantics* 1(1). 75–116.

Rooth, Mats. 1996. On the interface principles for intonational focus. In T. Galloway & J. Spence (eds.), *Proceedings of semantics and linguistic theory (salt) vi*, 202–226. Ithaca, NY: Cornell Working Papers in Linguistics (Cornell University). doi:10.3765/salt.v6i0.2767.

Rooth, Mats. 2009. Second occurrence focus and relativized stress F. In C. Fery M. Zimmerman (ed.), *Information structure:theoretical, typological, and experimental perspectives*, 15–35. Oxford: Oxford University Press.

Rooth, Mats. 2015. Representing focus scoping over new. In Thuy Bai & Deniz Özyildiz (eds.), *Proceedings of the forty-fifth annual meeting of the north east linguistic society.*, GLSA, Amherst, MA.

Rooth, Mats. 2016. Alternative semantics. In *The oxford handbook of information structure*, Oxford University Press.

Rooth, Mats, Jonathan Howell & Michael Wagner. 2013. Harvesting speech datasets for linguistic research on the web. *Final project white paper, Digging into Data Challenge* http://hdl.handle.net/1813/34477.

Sarojini, Balakrishnan, Narayanasamy Ramaraj & Savarimuthu Nickolas. 2009. Enhancing the performance of libsvm classifier by kernel f-score feature selection. In *Contemporary computing*, 533–543. Springer.

Saussure, Ferdinand de. 1967[1916]. *Cours de linguistique generale*. Harrassowitz.

Schwarzschild, Roger. 1999. Givenness, avoid f and other constraints on the placement of accent. *Natural Language Semantics* 7. 141–177. doi:10.1023/A:1008370902407.

Selkirk, Elisabeth. 1996. The prosodic structure of function words. In J.L. Morgan & K. Demuth (eds.), *Signal to syntax*, 187–213. Mahwah, NJ: Lawrence Erlbaum Associates.

Selkirk, Elisabeth O. 1984. *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Selkirk, Elizabeth O. 1995. Sentence prosody: Intonation, stress, and phrasing. In John A. Goldsmith (ed.), *Handbook of phonological theory*, chap. 16, 550–569. London: Blackwell.

Sluijter, A.M.C. & V.J. van Heuven. 1996. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 100. 2471–2485. doi:10.1121/1.417955.

Stevens, J. 2002. *Applied multivariate statistics for the social sciences*. Mahwah, NJ London: Lawrence Erlbaum.

Trubetzkoy, N.S. 1939. Grundzüge der Phonologie. *Travaux du cercle linguistique de Prague* 7.

Turk, Alice, Satsuki Nakai & Mariko Sugahara. 2006. Acoustic segment durations in prosodic research: A practical guide. In S. Sudhoff (ed.), *Methods in empirical prosody research*, vol. 3 Language Context and Cognition, 1–27. Berlin: Mouton de Gruyter. doi:10.1515/9783110914641.1.

Vanderslice, R. & P. Ladefoged. 1972. Binary Suprasegmental Features and Transformational Word-Accentuation Rules. *Language* 48(4). 819–838. doi:10.2307/411990.

Venables, W.N. & B.D. Ripley. 2002. *Modern applied statistics with s*. Springer-Verlag. doi:10.1007/978-0-387-21706-2.

Wagner, Michael. 2005. *Prosody and recursion*: MIT dissertation.

Wagner, Michael. 2006. Givenness and locality. In Masayuki Gibson & Jonathan Howell (eds.), *Proceedings of SALT XVI*, 295–312. Ithaca, NY: CLC Publications.

Weston, Jason, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio & Vladimir Vapnik. 2001. Feature selection for svms. In Volker Tresp Todd K. Leen, Thomas G. Dietterich (ed.), *Advances in neural information processing systems (nips)*, 668–674. San Mateo, CA: Morgan Kaufmann.

Williams, Edwin. 1997. Blocking and anaphora. *Linguistic Inquiry* 28(4). 577–628.

Winkler, Susanne. 1996. *Focus and secondary predication*. Berlin: Mouton de Gruyter.

Xu, Yi, Ching X Xu & Xuejing Sun. 2004. On the temporal domain of focus. In *Speech prosody 2004, international conference*, .

A Acoustic Measures

For each utterance of “than I did”, the following phonetic segments were annotated: 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*. Ratio refers to V2/V3.

Acoustic measure	Description
duration_{V1,V2,V3,C3,ratio}	duration of segment (vowel, stop closure)
meanIntensity_{V1,V2,V3,ratio}	RMS Intensity over vowel
maxIntensity_{V1,V2,V3,ratio}	max RMS Intensity
minIntensity_{V1,V2,V3}	min RMS Intensity
rangeIntensity_{V1,V2,V3}	range of RMS Intensity in vowel
maxIntTime_{V1,V2,V3}	time of intensity max relative to vowel duration
minIntTime_{V1,V2,V3, ratio}	time of intensity min relative to vowel duration
energy_{V1,V2,V3,ratio}	mean energy over vowel
power_{V1,V2,V3,ratio}	mean power of vowel
amp_{V1,V2,V3,ratio}	mean amplitude of vowel
pulses_{V1,V2,V3}	number of glottal pulses
jitter_{V1,V2,V3, ratio}	jitter
shimmer_{V1,V2,V3, ratio}	shimmer
f0_{V1,V2,V3, ratio}	mean F_0 of vowel
maxf0_{V1,V2,V3, ratio}	max F_0 of vowel
minf0_[V1,V2,V3, ratio}	min F_0 of vowel
maxF0Time_{V1,V2,V3, ratio}	time of F_0 max relative to vowel duration
minF0Time_{V1,V2,V3, ratio}	time of F_0 min relative to vowel duration
rangeF0_{V1,V2,V3, ratio}	f0max - f0min
maxf1_{V1,V2,V3, ratio}	max F1 of vowel
minf1_{V1,V2,V3, ratio}	min F1 of vowel
maxf1Time_{V1,V2,V3, ratio}	time of F1 max relative to vowel duration
minf1Time_{V1,V2,V3, ratio}	time of F1 min relative to vowel duration
rangef1_{V1,V2,V3, ratio}	F1max - F1min
f1TimeIntmax_{V1,V2,V3}	F1 value at time of intensity max
f1TimeF0max_{V1,V2,V3}	F1 value at time of F_0 maximum
f1Time{10,20... 90}_{V1,V2,V3}	F1 value at 10% 20%... 90% of vowel duration
f1bandIntmax_{V1,V2,V3}	F1 bandwidth value at time of intensity max

Acoustic measure	Description
f1bandF0max_{V1,V2,V3}	F1 bandwidth value at time of F_0maximum
f1band{10,20... 90}_{V1,V2,V3}	F1 bandwidth value at 10% 20%... 90% of vowel duration
maxf2_{V1,V2,V3, ratio}	max F2 of vowel
minf2_{V1,V2,V3, ratio}	min F2 of vowel
maxf2Time_{V1,V2,V3, ratio}	time of F2 max relative to vowel duration
minf2Time_{V1,V2,V3, ratio}	time of F2 min relative to vowel duration
rangef2_{V1,V2,V3, ratio}	F2max - F2min
f2TimeIntmax_{V1,V2,V3}	F2 value at time of intensity max
f2TimeF0max_{V1,V2,V3}	F2 value at time of F$_0$ maximum
f2Time{10,20... 90}_{V1,V2,V3}	F2 value at 10% 20%... 90% of vowel duration
f2bandIntmax_{V1,V2,V3}	F2 bandwidth value at time of intensity maximum
f2bandF0max_{V1,V2,V3}	F2 bandwidth value at time of F$_0$maximum
f2band{10,20... 90}_{V1,V2,V3}	F2 bandwidth value at 10% 20%... 90% of vowel duration
f1f2TimeIntmax_{V1,V2,V3}	F2-F1 at time of intensity max
f1f2Timef0max_{V1,V2,V3}	F2-F1 at time of F_0 max
f1f2Time{10,20... 90}_{V1,V2,V3}	F2-F1 value at 10% 20%... 90% of vowel duration
h1minush2p0_{V1,V2,V3}	1st harmonic minus 2nd harmonic at time of F_0 maximum
h1minush3p0_{V1,V2,V3}	1st harmonic minus 3rd harmonic at time of F_0 maximum
h2minush3p0_{V1,V2,V3}	2nd harmonic minus 3rd harmonic at time of F_0 maximum
h1minusa1p0_{V1,V2,V3}	1st harmonic minus amplitude of first formant at time of F_0 maximum
h1minusa2p0_{V1,V2,V3}	1st harmonic minus amplitude of second formant at time of F_0 maximum
h1minusa3p0_{V1,V2,V3}	1st harmonic minus amplitude of third formant at time of F_0 maximum
h1minush2f1_{V1,V2,V3}	1st harmonic minus 2nd harmonic at time of F1 maximum
h1minush3f1_{V1,V2,V3}	1st harmonic minus 3rd harmonic at time of F1 maximum
h2minush3f1_{V1,V2,V3}	2nd harmonic minus 3rd harmonic at time of F1 maximum
h1minusa1f1_{V1,V2,V3}	1st harmonic minus amplitude of first formant at time of F1 maximum
h1minusa2f1_{V1,V2,V3}	1st harmonic minus amplitude of second formant at time of F1 maximum

Acoustic measure	Description
h1minusa3f1_{V1,V2,V3]}	1st harmonic minus amplitude of third formant at time of F1 maximum

B Stimuli for Perception Experiment

Item	Sentence type	Focus category	Text
1	declarative	s	At first, you made a very small amount more than I did. Then after a year or two you made much more than I did.
2	declarative	s	I think Tom said it a little better than I did. In fact, he said it a lot better than I did.
3	declarative	ns	I'll feel probably 90% better than I did last week. In fact, maybe 100% better than I did.
4	declarative	ns	Today, I know a little bit more than I did when I started. And in a few weeks, I'll know way more than I did.
5	declarative	s	You worked harder than I did, and you worked longer than I did.
6	declarative	s	Tom knew more than I did, and he remembered more than I did.
7	declarative	ns	I feel generally more pessimistic now than I did as a kid, and I feel more conservative than I did as a kid, as well.
8	declarative	ns	I felt more comfortable onstage than I did offstage. And I felt more confident onstage than I did offstage.
9	declarative	s	There were a lot of photographers who would shoot more than I did.
10	declarative	s	He saw the situation differently than I did.
11	declarative	ns	I learned more in the last three hours than I did in the last three years of high school.
12	declarative	ns	I've been traveling more than I did when I was playing full time, so it's time to slow down.
13	interrogative	s	Why would anyone stay there longer than I did?
14	interrogative	s	How can I help my kids to achieve more than I did?

Item	Sentence type	Focus category	Text
15	interrogative	ns	Why do I have more energy today than I did the day before?
16	interrogative	ns	How can I find time to visit my family this year more than I did last year?

C Feature sets selected by Boruta algorithm for lab dataset

Features selected by Boruta algorithm from full feature set

duration_V2
duration_V3
pulses_V2
pulses_V3
jitter_V2
jitter_V3
shimmer_V3
f0_ratio
maxf0_ratio
minf0_V3
minf0_ratio
minf0Time_V3
minf0Time_ratio
rangef0_V3
meanIntensity_V3
meanIntensity_ratio
maxIntensity_V3
maxIntensity_ratio
minIntensity_V3
minIntensity_ratio
maxIntTime_V3_percent
maxIntTime_ratio
minIntTime_V3_percent
minIntTime_ratio
energy_ratio
power_V3
power_ratio
f1Time10_V2
maxf1_V2
rangef1_V2
f1TimeIntmax_V2
f2Time10_V2
minf2_V2
[f2TimeIntmax_V2
f1f2Time10_V2
f1f2TimeIntmax_V2
f1Time10_V3
f1Time90_V3
minf1_V3
f1bandTime10_V3
f2bandTime10_V3
f1Time20_V2
f2Time20_V2
f1Time30_V2
f2Time30_V2
f1Time40_V2
f2Time40_V2
f1Time50_V2
f2Time50_V2
f1Time60_V2
f2Time60_V2
f1Time70_V2
f1Time80_V2
f1bandTime20_V3
f1Time30_V3
f1Time40_V3
f1Time50_V3
f1Time60_V3
f1Time70_V3
f1Time80_V3
f1f2Time20_V2
f1f2Time30_V2
f1f2Time40_V2
f1f2Time50_V2
f1f2Time60_V2
f1f2Time70_V2

Features selected by Boruta algorithm from set of *f0* features

f0_V2
f0_V3
f0_ratio
maxf0_V2 maxf0_V3
maxf0_ratio
minf0_V2
minf0_V3
minf0_ratio
maxf0Time_V3
maxf0Time_ratio
minf0Time_V3
minf0Time_ratio
rangef0_V2
rangef0_V3
rangef0_ratio
f0_V1
maxf0_V1
minf0_V1

Features selected by Boruta algorithm from set of non-f0 features		
duration_V2	power_ratio	f2Time40_V2
duration_V3	amp_ratio	f1Time50_V2
pulses_V2	f1Time10_V2	f2Time50_V2
pulses_V3	maxf1_V2	f1Time60_V2
pulses_ratio	rangef1_V2	f2Time60_V2
jitter_V2	f1TimeIntmax_V2	f1Time70_V2
jitter_V3	f2Time10_V2	f1Time80_V2
shimmer_V3	minf2_V2	f1bandTime20_V3
meanIntensity_V3	f2TimeIntmax_V2	f1Time30_V3
meanIntensity_ratio	f1f2Time10_V2	f1Time40_V3
maxIntensity_V3	f1f2TimeIntmax_V2	f1Time50_V3
maxIntensity_ratio	minf1_V3	f1Time60_V3
minIntensity_V3	f1Timef0max_V3	f1Time70_V3
minIntensity_ratio	f1bandTime10_V3	f1Time80_V3
maxIntTime_V3	f2bandTime10_V3	f1f2Time20_V2
maxIntTime_ratio	f1Time20_V2	f1f2Time30_V2
minIntTime_V3	f2Time20_V2	f1f2Time40_V2
minIntTime_ratio	f1Time30_V2	f1f2Time50_V2
energy_ratio	f2Time30_V2	f1f2Time60_V2
power_V3	f1Time40_V2	f1f2Time70_V2

Features selected by Boruta algorithm from set of syntagmatic features		
pulses_ratio	rangef0_ratio	minIntTime_ratio
f0_ratio	meanIntensity_ratio	energy_ratio
maxf0_ratio	maxIntensity_ratio	power_ratio
minf0_ratio	minIntensity_ratio	amp_ratio
minf0Time_ratio	maxIntTime_ratio	duration_ratio

Features selected by Boruta algorithm from set of paradigmatic features

duration_V3
pulses_V2
pulses_V3
jitter_V2
jitter_V3
shimmer_V3
f0_V3
minf0_V3
minf0Time_V3
rangef0_V3
meanIntensity_V3
maxIntensity_V3
minIntensity_V3
maxIntTime_V3
minIntTime_V3
power_V3
f1Time10_V2
maxf1_V2
rangef1_V2
f1TimeIntmax_V2
f2Time10_V2
minf2_V2
f2TimeIntmax_V2
f1f2Time10_V2
f1f2TimeIntmax_V2
f1Time10_V3
f1Time90_V3
minf1_V3
f1Timef0max_V3
f1bandTime10_V3
f2bandTime10_V3
f1f2Time10_V3
f1Time20_V2
f2Time20_V2
f1Time30_V2
f2Time30_V2
f1Time40_V2
f2Time40_V2
f1Time50_V2
f2Time50_V2
f1Time60_V2
f2Time60_V2
f1Time70_V2
f2Time70_V2
f1Time80_V2
f1Time20_V3
f1bandTime20_V3
f1Time30_V3
f1Time40_V3
f1Time50_V3
f1Time60_V3
f1Time70_V3
f1Time80_V3
f1f2Time20_V2
f1f2Time30_V2
f1f2Time40_V2
f1f2Time50_V2
f1f2Time60_V2
f1f2Time70_V2