Non-Native Differences in Prosodic-Construction Use

Nigel G. Ward
Department of Computer Science
University of Texas at El Paso and Kyoto University
El Paso, TX, 79928, USA

Paola Gallardo
Department of Computer Science
University of Texas at El Paso
El Paso, TX, 79928, USA

Editor: Dr. Gregory Aist

Note to Editor: This is a revised version of our submission, modified because further explorations with a new method gave new findings, and thus, instead of a negative result overall, the paper now reports that we did find significant differences.

Compared to the original (August 2015), the major change is the new Section 7 on the new method and results, and a mostly new conclusion (Section 8). It’s also been restructured slightly: the Data section comes earlier, old Sections 5 and 7 are combined into new Section 6, and the old section on Spanish prosody is gone, with the relevant bits moved elsewhere. There are other minor improvements, marked with changebars.

Also, just as information, we’ve submitted a paper on the new method to Interspeech 2016, although depending on the outcome of this submission we may withdraw it.

Abstract

Many language learners never acquire truly native-sounding prosody. Previous work has suggested that this involves skill deficits in the dialog-related uses of prosody, and may be attributable to weaknesses with specific prosodic constructions, but this has not been systematically investigated. Using semi-automated methods, we identified 32 of the most common prosodic constructions in English dialog. We performed the same analysis for 90 minutes of six advanced native-Spanish learners conversing in English, and we then compared the two sets. Three prosodic constructions common in the English-native data were saliently absent from the learners data, relating to the pragmatic functions of showing involvement, showing lack of involvement, and explaining from a factual perspective.

Keywords: conversation, dialog, interaction, pragmatics, language learning, non-native prosody, second language, L2, American English, Mexican Spanish speakers, L1 transfer, cross-cultural communication, misperceptions, turn-taking
1. Introduction

Non-native speakers often have saliently non-native prosody, even when their other language skills are good (Zimmerer et al., 2014). Among the various functions of prosody, it has been suggested that the dialog-related aspects might be most important for language learners, since incomplete command of the prosodic forms used for pragmatic functions can impact interactional competence and the achievement of communicative goals (Barraja-Rohan, 2011). Further, non-natives may show an “over-use of a limited variety of intonation patterns in the L2” and an “underuse” of others (Ramirez Verdugo, 2003, 2006).

This paper reports a corpus-based study of these issues. Specifically, we examine the extent to which non-native dialog-related prosody (“dialog prosody”) differs from that of native speakers, and whether there are deficits with specific prosodic constructions.

This paper is organized as follows. First we discuss previous approaches and their limitations (Section 2). Choosing to describe prosodic skills in terms of prosodic constructions, we applied Principal Components Analysis (PCA) (Section 3) to native-native dialog data, resulting in the identification of 32 common prosodic constructions of English dialog (Section 4). No suitable data collections existing, we collected 90 minutes of data from six advanced native-Spanish speakers in conversations with native speakers (Section 5). Simple statistical measures revealed some differences in construction use (Section 6), but it was more informative to do comparisons at the level of models. Thus we first we found the non-natives’ prosodic constructions, and then compared these to the natives constructions; this revealed specific missing skills (Section 7). In Section 8 we summarize and note questions for future research.

2. Methods for Characterizing Non-Native Prosodic Differences

The problem of characterizing non-native prosody and how it differs from native-speaker prosody is complex. This is especially true for the dialog-related uses of prosody. This section overviews four commonly used approaches and notes their limitations.

The first approach starts with a pragmatic function. The classic example is Gumperz’s discussion of cafeteria servers who offered side dishes using a falling accent, a function which English native speakers perform with a rising accent (Gumperz, 1982). Other work in this vein has examined English question, focus and list-final intonation (Ramirez Verdugo, 2003; Swerts and Zerbian, 2010; Kainada and Lengeris, 2015), Italian contrast (Turco et al., 2015), and Spanish turn keeping and information seeking (Aronsson and Fant, 2014). Such analyses rely on the existence of a clear intended function, and thus they cannot give us the whole story, since speakers in dialog commonly pursue multiple goals with each utterance (Bunt, 2011; Heritage, 2012). In such cases it is impossible to say definitively what the appropriate prosodic form would be, as native speakers in the same situation might differ in which of the pragmatic functions they choose to prosodically highlight.

A second approach starts with a syntactic form or sentence type and looks at differences between learners’ and natives’ prosodic realizations of it. While often informative, this approach is complicated by the fact that a given syntactic structure may serve different discourse functions, depending on the context and the prosody. Indeed, as prosody is often determined more by the pragmatic function than the syntactic form, especially in dialog (Lai, 2012; Hedberg
et al., 2014), findings obtained from such production tasks may not fully reflect actual behavior in dialog. Nevertheless most detailed studies of learners’ prosody have used this approach.

A third approach characterizes non-native prosody with reference to a model of the appropriate prosodic forms. For example, Toivanen’s examination of the distribution of tone types (fall, rise-fall, fall-plus-rise, etc.) showed that learners used rising tones less than did native speakers (Toivanen, 2003). However model-based analysis also has its limitations. For models of prosody that are symbolic rather than phonetic, labor-intensive segmentation and/or hand labeling is required before they can be applied to data. More generally, such approaches only work for the aspects of prosody that a model handles, but of course these are always limited. For example, the most popular current models of prosodic forms are based on only monolog data, and mostly handle only pitch (intonation), and they leave out speaking rate, timing, and energy information, but these aspects are of course important in modeling learners’ skills (Trouvain and Gut, 2007; Romero-Trillo, 2012).

A fourth approach uses raw statistics on prosodic usage over corpora. For example, Zimmerer’s measurements of native-speaker and non-native corpora show much less pitch variation in the latter (Zimmerer et al., 2014). This method can exploit large amounts of data and is entirely objective. It is also robust: for example, while in any given utterance a non-native may have a good reason to use compressed pitch range — for example when losing interest in a topic and preparing to close it out — consistent use of less pitch variation than native speakers is good evidence for something missing. Other work in this vein has shown that values of speaking rate and pitch range correlate with assessments of comprehensibility and accentedness (Kang, 2010). However raw statistics, being context-independent, cannot pinpoint the locations of the differences, nor their communicative significance. For example, they cannot tell in which specific contexts more pitch range would have been appropriate.

These methods have provided insights into the issues, and details regarding some specific differences in prosody. Nevertheless, there is as yet no big-picture understanding of non-natives’ dialog-prosodic skills. In particular, we want to know whether it is indeed in dialog-related functions that their prosody differs most, and what specific dialog-related skills they are weakest at.

Our method for investigating these questions, like the latter two approaches above, starts with forms, rather than with functions, and is corpus-based. This is because we are centrally interested in what people actually do in conversation, taking inspiration from the idea that the pragmatic functions that matter most in real life are those which people actually use most in conversation.

3. Prosodic Constructions and their Automatic Discovery

Describing prosodic behavior is difficult. There is currently no consensus on how to represent prosodic knowledge, and prosody as used for dialog purposes is especially problematic (Kalathottukaren et al., 2015). For this study we chose to use an approach based on an inventory of constructions, since this can support both automated analysis and analysis with respect to pragmatic functions and dialog skills.
### Table 1: Major Components of a Prototypical Rendition of the Upgraded Assessment Construction. Times are in milliseconds relative to the end of Speaker A’s assessment.

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker A</th>
<th>Speaker B</th>
</tr>
</thead>
<tbody>
<tr>
<td>–3000 to –300ms</td>
<td>speaking</td>
<td>quiet</td>
</tr>
<tr>
<td>–2500 to –1300</td>
<td>speaking louder</td>
<td>quiet</td>
</tr>
<tr>
<td>–800 to 0</td>
<td>tapering in loudness to silence</td>
<td></td>
</tr>
<tr>
<td>–300 to +300</td>
<td>quiet or silent</td>
<td>loud and fast</td>
</tr>
<tr>
<td>+300 to +600</td>
<td>resuming speaking</td>
<td>slowing and fading out</td>
</tr>
<tr>
<td>+600 to +1300</td>
<td>speaking fast</td>
<td>quiet</td>
</tr>
<tr>
<td>+1300 to +3200</td>
<td>speaking</td>
<td>quiet</td>
</tr>
</tbody>
</table>

3.1 Prosodic Constructions

Recently a shared notion of prosodic construction has emerged from work in several research traditions, including conversation analysis, experimental phonetics, autosegmental-metrical intonation modeling, and big-data analysis (Ogden, 2007, 2012; Petrone and Niebuhr, 2013; Niebuhr, 2014; Hedberg et al., 2014; Ward, 2014). Prosodic constructions are recurring temporal patterns of prosodic activity that express specific meanings and functions. They typically involve not only pitch contours but also energy, rate, timing and articulation properties, and may involve synchronized contributions by two participants.

For example, in the Upgraded Assessment Construction, as described by Ogden (Ogden, 2012; Ward, 2014), a listener expresses agreement with an assessment by producing an upgraded version, for example when one speaker (A) observes it’s pretty and the other (B) follows with absolutely gorgeous. The upgraded assessment is generally produced with increased amplitude, pitch height, pitch range, and rate, and with a ‘tighter’ articulation.

Often this upgraded assessment follows a bid for some kind of empathy or affiliation. Prosodically this involves A speaking loudly for a bit but then trailing off, where the trailing-off is in a lower pitch, and then falling silent for a moment. B’s upgraded assessment in turn is often followed by resumed speech by A that is again louder and tends to last for a few seconds.

Thus this construction involves interleaved prosodic behaviors by two participants, with specific sequencing and timing. Table 1 roughly shows the prototypical temporal configuration of this construction. In jointly performing this construction the participants each express specific attitudes, and together establish a shared assessment and joint interest.

Prosodic constructions share much with the classical notion of intonation contour (Liberman and Sag, 1974; Ladd, 1978). They describe a recurring sequence of prosodic elements in a specific temporal configuration, with some dialog function. These functions often affect the future course of the dialog or the unfolding relationship between the participants. They may in addition have meanings or expressive values, although these are often abstract and highly context-dependent.

Prosodic constructions extend intonation contours in three ways (Niebuhr, 2014; Ward, 2014). First, they describe not only patterns of pitch but also of other prosodic features, such as amplitude, rate, and timing. Second, prosodic constructions are not limited to the behavior of a
single speaker, but often describe coordinated actions by two parties. Third, they are not necessarily linked to sentences or utterances, but instead can cover arbitrary regions of time. Prosodic constructions resemble grammatical constructions (Goldberg, 2013) — form-function pairings where the form is a syntactic template and the function is some conventionalized semantic or pragmatic content — in particular in being composable.

Constructions can be modeled in various ways: qualitatively, symbolically (Hedberg et al., 2014), or quantitatively (Lai, 2012). For this paper we use quantitative descriptions, as they have two useful properties. First they are superimposable, which suits the fact that any specific time in a dialog may involve multiple prosodic behavior sequences expressing simultaneously present pragmatic functions. Second, their presence is graded, meaning that a construction is not simply present or absent, rather it can be present to varying degrees, to the extent that more of the component features are more strongly present and their temporal configuration more closely matches the prototype. For example, a weak version of the upgraded assessment construction might function as a somewhat perfunctory acknowledgment.

Despite some limitations (Ward, 2014), this approach to prosody has the important advantage of enabling the automatic detection of prosodic constructions in unlabeled data. This enables the computation of statistics on construction use.

3.2 Construction Discovery by Principal Components Analysis

In order to examine non-natives’ uses of dialog prosody as comprehensively as possible, we need a large inventory of constructions.

Constructions can be discovered in many ways, including inductively by conversation analysis (Ogden, 2012), statistically by corpus-based studies of the realizations of known pragmatic functions (Hedberg et al., 2003; Niebuhr, 2014; Hedberg et al., 2014), and semi-automatically (Ward, 2014). Given similar data, it appears that these methods can all give similar results, so here we used the fastest and easiest: a semi-automated one.

Several automated and semi-automated discovery methods for intonation contours and other prosodic elements have recently been developed. Techniques include clustering, Functional Data Analysis (FDA), and Principal Components Analysis (PCA) (Jokisch et al., 2014; Reichel, 2014). In this paper we use PCA, because it is relatively simple, and because it works for raw dialog data, without needing preliminary segmentation or annotation.

PCA can be described in several ways, but it is convenient to view it as an iterative analysis process. In each stage, PCA finds the factor that explains as much as possible of the observed variation, across many datapoints and many variables. It then subtracts out what that factor explains, finds another factor to explain much of the remaining variation, and iterates. For example, if we have statistics on children, including height, weight, running speed, arm strength, lung capacity, stamina, and so on, the first underlying factor may be age, the second something like skinny-chubby, the third socioeconomic status, and so on. The observed variable values for any datapoint (child) are modeled as linear combinations of these underlying factors. Conversely, given the observed values for a datapoint, it is trivial to compute the values of the underlying factors, by a simple matrix multiplication.

Prosodic constructions as we model them — being graded and superimposable — perfectly suit the assumptions of PCA: they can serve as the underlying factors that explain the surface,
observed, prosody. That is, the observed prosody over any short region of a dialog can be explained as the superimposed effects of multiple, simultaneously-active constructions. Thus, our method is to apply PCA to datapoints, each of which is a point in time, each described by various observed prosodic features. The output is then a set of dimensions, which are configurations of features that frequently occur together.

### 3.3 Prosodic Features Used

This subsection documents the prosodic features used as input to PCA. We used a set of features and windows chosen to be relevant to the dialog-related-aspects of prosody, rather than lexical, syntactic, emotional, or speaker-dependent aspects. In particular, this set includes features of both speakers’ prosodic behavior, since we are interested in constructions involving contributions by both speakers. Furthermore, our set is computed over fixed-width windows at fixed offsets, rather than using syllable-, word-, or utterance-aligned features. This is because, first, although useful for many purposes (including studies of speaker differences, lexical-accent realizations, and speech-act forms), aligned features seem relatively less relevant to dialog behaviors; second, aligned features are impossible to accurately compute automatically, and third it is difficult to even define aligned features when considering with both sides of a dialog, since speakers seldom behave in lockstep.

To support the discovery of temporal patterns, for each datapoint we use a number of features at different offsets to broadly represent the local prosodic context. For example, in addition to the volume over the past 50 milliseconds, we also use the volume over a 50 millisecond window centered 75 ms in the past, over a 100 ms window centered 150 ms in the past, and so on, for both past and future windows, spanning about 6 seconds centered around the point of interest. Including such offset features enables the use of PCA for time-series analysis. Specifically, for each datapoint we compute a set of 176 features, as listed in Figure 1. This resembles other recent prosodic feature sets used for machine-learning applications; in particular it includes not only pitch but also features for speaking rate, volume, and creaky voice (Schuller, 2011; Shriberg and Stolcke, 2004; Ward et al., 2011). We followed previous work in using more windows for the most informative features (notably amplitude), and in choosing the window sizes to give greater temporal resolution near the point of interest. Our features were computed using the Mid-Level toolkit (http://www.cs.utep.edu/nigel/midlevel/), using the built-in normalizations to make the output features fairly speaker-independent.

For completeness, we note that here, as always with prosodic features, normalization is an issue. The Mid-Level Toolkit outputs simple but reasonably robust methods approximately. For loudness we use log energy normalized per track to correct for different recording conditions and different speakers. For the pitch-height and pitch-range features we use percentiles in the distribution of pitch seen for that track, thus again normalizing for speaker. For rate, we use a simple frame-by-frame energy-difference measure. To avoid the problems associated with interpolating pitch over nonvoiced regions, we use as evidence for the associated pitch features (low, high, narrow wide) only the valid pitch points.
Figure 1: The Prosodic Feature Inventory. Start and end times for each window is in milliseconds offset from the point of interest. These features are computed for both left and right speakers, giving 176 in total.

### 3.4 From Dimensions to Constructions

While PCA is merely a mathematical operation, it often succeeds in identifying meaningful factors (dimensions) that underlie the observed behavior, and this has been seen for prosody also (Ward and Vega, 2012; Ward, 2014).

The workflow is summarized in Figure 2. For each timepoint in the corpus the prosodic features are computed. Principal Component Analysis digests all this data and outputs 176 dimensions. Each dimension has a weight on each of the features. For example, on the data set discussed below, Dimension 1 (principal component 1) has a high negative weight on Speaker-A-amplitude-over-0-50-milliseconds, a high negative weight on Speaker-A-amplitude-over-50-100-milliseconds, a positive weight on Speaker-B-energy-amplitude-over-0-50-milliseconds, and so on. Thus, at times when Speaker A is speaking and B silent, the value on dimension 1 will be negative, and for the opposite positive. Thus every dimension incorporates, in effect, two patterns.

Patterns often have temporal variation in the loadings. For example, for Dimension 1 the loadings on the high-pitch features are high for early windows but then fall, to the extent that, by the 800-1600 millisecond window, the low-pitch loading is greater than the high-pitch loading. Thus, the simple process of PCA finds the well-known prosodic phenomenon of declination. All patterns observed so far involve such temporal variation, so it is appropriate to call them constructions.
4. Some Prosodic Constructions of English

To quantify non-native behavior patterns, we need to compare them to some standard. Specifically, to do construction-based analysis we needed a reference data set to use for construction discovery. Wanting something representative of the language norm that our non-natives would be most familiar with, we decided to use the Social Speech collection (Ward and Werner, 2013). Like the primary data set, described below, this consists of unconstrained conversations among computer science students at a university in the Southwest United States, although recorded for a different purpose, recorded two years earlier, and recorded with different microphones. We took 6 native-native conversations from this corpus, lasting about 10 minutes each, and computed the prosodic features described above for 720,000 data points, taken every 10 milliseconds for both speakers. We then applied PCA.

The second column of Table 2 shows the percentage of variance accounted for by each of the resulting dimensions. It shows, for example, that describing prosodic behavior just with one value, the value on Dimension 1, explains 17% of the variance across all 176 features. Together the top 16 dimensions account for 55% percent of the variation in these dialogs, suggesting that examination of the top 32 constructions can cover most of the dialog-prosody skillset that learners need.

The other outcome of PCA was the prosodic description of each construction: the actual weightings, for each component, of each of the features. For example, Dimension 1 had a loading of –0.08 on the speaker A amplitude-over-800-to-1600-ms feature. As each dimension has a loading on each of the 176 features, it is convenient to use visualizations. For example, Figure 3 shows the loadings for Dimension 3. It can be seen that the loading for the speaker-A-log-energy (“volume”) feature from -1600 to -800 ms is positive, and so on. Examining the other loadings, it is clear that this dimension involves the A speaker (top) speaking and then falling silent, and the B speaker, conversely, being silent and then speaking. Thus it encompasses a turn-yielding construction (“dimension 3 lo”) and a turn-taking construction (“dimension 3 hi”). From the figure it is easy to also see some of the prosodic correlates typical of turn-yielding pattern in English: notably increases in volume, speaking rate, and creakiness, followed by a further increases on the latter two and a simultaneous drop in pitch. Of course, specific instances of turn yield will not follow this pattern exactly, due to the simultaneous presence of other, superimposed, constructions. An unusually well-matching ex-
Figure 3: Loadings of Dimension 3. Purple solid lines are for the A speaker; green dashed lines for B. Time is in milliseconds. The dotted lines are zeros, with points above them indicating positively loaded features and points below negative. The “pitch height” line shows the difference between the loadings of the high-pitch and low-pitch features; similarly “pitch width” is the difference of the wide and narrow features. While this figure shows the strengths of factor loadings, rather than average values for pitch height etc., in practice, instances in the dialogs where this dimension is strongly present do tend to have feature values varying over time as this visualization suggests. While the volume features extend out to 3200 ms before and after the point of interest, to save space we show only 4 seconds-worth of feature loadings.

Example is seen in Figure 4, transcribed as Example 1. Audio for the examples is provided at http://www.cs.utep.edu/nigel/12english/.

We note in passing that it is interesting that this method assigns the pitch rise to the same construction as the swift turn exchange, based on correlations. Other analyses might consider it instead a nuclear accent, not directly related to turn-taking. Determining which of these accounts, if either, best matches psychological reality would be an interesting question to pursue. For purposes of analyzing non-native differences, however, we proceed without resolving such questions.

It is perhaps worth also noting that in Dimension 3 the loadings of features for the two speakers are symmetric: past-future mirror images across the point of interest (0 milliseconds). We do not ascribe any deep significance to this: PCA often results in dimensions with some form of symmetry, and this tendency is stronger here, because the features computed for the
Figure 4: A Swift Turn Exchange, high on Dimension 3. Pitch is shown at the bottom of each track.

two sides are identical, and because the two sides are slightly correlated, due to a small amount of cross-track bleeding.

For reasons of space we below only discuss the loadings when relevant to non-native differences. (All the loadings are available at http://www.cs.utep.edu/nigel/l2english/, both numerically and as visualizations.)

Example 1 (soc008@165.1)

A: I just need to get that lab done, and I’m done with that lab.

B: What, what about, where, where did you guys get in the homework?

While Dimension 3 was easy to understand, this was not always true. To interpret each dimension we applied an eclectic mix of methods. We considered feature loadings in relation to information in the literature on the pragmatic functions of the strongly-loaded features. We also examined places in the corpus where a construction was strongly present. We used qualitative-inductive methods to find commonalities among such places, considering different pragmatic aspects, as revealed by the behaviors of both participants in the immediate context,
and occasionally also engaging our own intuitions of what was being conveyed by the observed prosodic form, in contrast to other forms that might have appeared. For some dimensions the commonalities were obvious, both from the features and the examples examined. For others the commonality did not become clear until we had examined many examples, and then another dozen or so to confirm the general tendency.

The interpretation process was complicated for several reasons. One is that the pragmatic force of any individual construction depends on the local context, including other constructions simultaneously present. For example, the swift turn exchange seen in Example 1, is not only high on Dimension 3, but also fairly high on Dimension 2, since there is a lot of talk by both speakers, and low on dimension 16, primarily since the last syllable of the bottom speaker is short and creaky, indicating his attitude towards the lab. Another reason interpretation was complicated was individual differences in behavior and uses. For example, at some of the times when one speaker’s prosody was high on Dimension 10, it seemed he was being provocative, somewhat different from the more common use, of disagreeing or diverging, although not unrelated. More generally, the constructions appear polysemous, for example, when low on Dimension 10, one speaker was often producing a short check questions, and/or saying something they expected the other to agree with, and the other usually did. In the table we generalize this as speakers “aligning,” but for other dimensions we sometimes list related functions rather than generalizing over them. Other analysts could make other choices. An additional complication for the analysis was the presence of creative and deliberate uses of prosody, including uses for non-literal meanings and for reported speech (Rao, 2013b; Estelles-Arguedas, in press, 2015), many of which did not follow the general tendencies.

Our interpretations are summarized in the third column in Table 2. Again these must be considered only tendencies: not every time point in the corpus prosodically fitting a listed construction also exhibits the listed function or meaning, for the reasons noted above. Each of these really deserves a paper in itself, to treat its form and function in detail, and to relate it to alternate ways of describing the phenomena. However, for reasons of space, we discuss further only those which turn out to be used differently by non-natives or otherwise interesting.

Ideally we would like a full and final listing of the prosodic constructions of English before going on to examine the non-native differences. Of course this listing falls short. In addition to the issues noted above, the details of the feature set we chose are somewhat arbitrary, and we have observed that with different feature sets the resulting dimensions vary slightly. Further, our method did not identify exclusively dialog-related aspects of prosody, nor, certainly, all of the dialog-relevant constructions, not least because our features only cover six-second spans. Thus we do not propose this listing as a universally valid or objectively-verified list of the pragmatic functions of English prosody. Nevertheless all of these functions have been previously identified in the literature as important for dialog (Wells, 2006; International Standards Organization, 2012; Riggenbach, 1991; Couper-Kuhlen and Selting, 1996; Sidnell, 2011; Clark, 1996; Reed, 2010), and PCA-based studies of other corpora have revealed similar constructions and functions (Ward and Vega, 2012; Ward, 2014), so this list does seem likely to be useful. For example, it may be useful for English-language curriculum design (Diepenbroek and Derwing, 2013; Busa, 2012). Below we use it for the analysis of non-native behaviors.

5. Data
For our non-native data we chose to record advanced non-native speakers. This choice was inspired by reports of those who, despite years of immersion, still have weak prosodic skills (Zimmerer et al., 2014), and from personal observation of friends and family members for which this is the case. In this we diverge from the common practice of studying non-native prosody using data from learners still in language classes. This section summarizes some of the important properties of our data sets; the details appear elsewhere (Ward and Gallardo, 2015).

We chose to record non-natives who had completed at least one semester of college in the United States. We excluded those with significant non-classroom English-language experience before age 17 or who otherwise seemed to have almost-native conversation skills. The corpus speakers had strong vocabulary and good fluency, but all were noticeably non-native in pronunciation. All had grown up in Northern Mexico and had Spanish as their native-language.

We chose native speakers of Spanish based on convenience. The segmental, lexical, and syntactic aspects of Spanish prosody are known to differ from English, as are some the expressions of some pragmatic functions, including questions, back-channeling, complaining, and expressing probability and usuality (Bowen, 1956; Farias, 2013; Hualde, 2005; Berry, 1994; Ramirez Verdugo, 2005; Rivera and Ward, 2006; Rao, 2013a; Santiago and Delais-Roussarie, 2015; de la Mota et al., 2010). Spanish also expresses some pragmatic functions less with prosody than with word order, discourse particles, or gesture (Borras-Comes et al., 2014; Ortega-Llebaria and Colanton, 2014). Accordingly it seemed likely that there would be differences in dialog prosody also.

In recording we did not ask the speakers to do anything more specific than talk to each other. Their conversations were spontaneous and vary widely in topic. While there are advantages to using conversation data based on scripted or role-play interactions, spontaneous conversations may more closely approximate real-world interaction. While producing appropriate intonation in monolog or scripted dialog is, in essence, “merely” a question of choosing the appropriate intonation contour and applying it to a sentence, producing appropriate prosody in dialog is a much greater challenge. Realization of each construction requires using multiple prosodic features in specific temporal configurations. Moreover multiple constructions must often be simultaneously realized. This all must be done under the pressure of choosing words and listening to and coordinating with the dialog partner.

Each non-native was recorded in dialog with a native-speaker partner. We obtained 9 conversations, of about 10 minutes each, including 6 different non-native speakers.

We selected the non-natives for the corpus based on our perceptions of some degree of awkwardness with English, but without explicit consideration of prosodic behaviors. However we did a post-hoc examination to see whether there were, in fact, any non-native aspects to their prosody, regardless of whether these related to any specific construction or pragmatic function. There were indeed such differences: their prosody was non-native in many ways, most saliently in having: a tendency to syllable-timing rather than stress timing, unusual patterns of utterance-final lengthening or lack of lengthening, and misplaced stresses and accents. There seemed to be other differences but we did not attempt to categorize them, preferring to move directly to the model-based analyses.

It’s worth at this point noting two potential issues with this data. One is that, since each pair of speakers includes one native speaker, and since each pattern involves behavior by both speakers, it is possible that some observed differences could be due to natives speaking differently
when interacting with non-natives, rather than to differences in the behavior of the non-natives themselves. However we saw only rare evidence for this, and only for one speaker pair, so this is probably not a major problem. Another potential issue is that, statistically, a pattern may be detected as often used, when in fact this may be mostly due to times when the native speaker perfectly executes his side of the pattern, with little or no support from the non-native. Thus our method may understate the non-native differences.

In addition to the primary collection, of non-native speakers talking with monolingual English speakers, we recorded two other data sets: one of monolingual native English speakers talking with other native speakers, and one of Spanish speakers speaking together in Spanish. Both of these collections included many speakers from the primary collection. All were recorded in the same environment with the same equipment.

Finally, to test the validity of the first approach, we used another data set, the well-known Switchboard corpus (Godfrey et al., 1992). Table 3 summarizes the five data sets used.

6. Approach 1: Comparing Construction-Use Distributions

Expecting the dialog-prosodic deficit of non-natives to be largely associated with specific constructions, we set out to determine which.

In our first approach we computed statistics, across all the non-native data and timepoints, on the usage of each construction, to identify differences between the natives and the non-natives. Figure 5 overviews this workflow.

Given the dimensions, the prosody in the immediate context of every point can be represented as the sum of the contributions of all the dimensions active at that time; typically some positively and some negatively. Thus we applied the loadings discovered by PCA to samples taken every 10 milliseconds throughout the data. This was fully automatic; computationally just the dot product.
To test this method, we first applied it to another corpus of English, to see whether it would
find differences that made sense. Specifically we used 7 dialogs (14-speakers, 35 minutes total)
from Switchboard. Although also dyadic conversations in American English, in these conversa-
tions the participants were strangers, they were generally much older, they spoke by telephone,
and they started with suggested topics, such as crime and childcare, although most of the con-
versations rapidly moved on to other topics.

On Dimension 2 there was a large distribution difference relative to the reference data,
as seen in Figure 6: the Switchboard speakers exhibit fewer high values on this dimension,
meaning that less often were both participants are talking or laughing simultaneously. This can
be readily observed, and is unsurprising, given that more formal turn taking is generally found
in telephone conversations and in conversations between strangers.

For reasons of space we do not show the other distributions, but the columns 2 and 3 of
Table 4 show the means and standard deviations of Switchboard speakers’ uses of the top 16
dimensions, plus 2 more. The mean for the reference set is zero on each dimension, due to
normalization. Both the means and the standard deviations shown have been normalized by
(divided by) the standard deviation of same dimension in the reference set. Thus the units for
the means are standard deviations, with negative values where the Switchboard speakers tended
to be lower on that dimension and positive values when higher. For the standard deviations,
values less than 1 mean the Switchboard speakers had narrower distributions than the reference
speakers, and greater values wider distributions. We noted several differences.

For Dimension 10, the Switchboard speakers tend to the negative side. According to our
interpretations, this means that they exhibit more alignment and agreement than the reference
dialogs. This is again readily observable and unsurprising: strangers who have no desired out-
comes beyond having a pleasant conversation tend to find things they can agree on. Prosodically,
the 10-lo construction involves quieter-than-average utterances, with gradually increasing pitch
and occasional moments of faster speech with expanded pitch range.

Figure 6: Distribution of Values on Dimension 2 for the reference data and the Switchboard
data.
Dimension 14 also shows a large difference, tending to the low side, meaning less talk about mutually-known third parties and more about personal situations. This may relate to the lack of shared context and to a tendency to self-disclosure in a safe context as here when talking anonymously. Prosodically, the 14-low construction involves choppy short utterances with very brief pauses.

For Dimension 15, the Switchboard dialogs tend to higher values, expressing negative feelings about someone or something distant. This is again readily observable and unsurprising, since many of these conversations touch on issues like crime, childcare, taxes, and schools, and involve complaining about politics and institutions. This construction is complex, involving a sharp decrease in speaking rate and pitch range, resulting in a region of narrow pitch for about a second, for some speakers giving a muttering effect.

These examples indicate that this method can reveal differences in prosodic behavior that reflect real differences in dialog activities and interaction styles.

Having verified that examining distributions could discover meaningful differences, we applied the same method to the data for the nonnatives, and, for comparison, the monolingual native data.

Columns 4 through 7 of Table 4 show the means and standard deviations on each dimension for both sets. Most relevant are dimensions where the non-native behaviors differ not only from the reference data, but also from the native-speaker subset of the corpus. While we expect variation between any two random sets of speakers, if the non-natives differ from both the native data and the reference data, that suggests a real difference.

Comparing the means (columns 4 and 6 of the table), we first note the lack of striking differences: unexpectedly, the non-native means were mostly closer to the native means than were the Switchboard speakers’ means. We examined the differences statistically, using unmatched, two-tailed heteroskedastic t-tests with Bonferroni corrections. When we took as independent samples the means of each speaker’s values there were no statistically significant differences, doubtless due to the small number of speakers. When instead we took the speakers’ means over 30-second samples as independent, some were, as shown by the asterisks in column 6 of the table. The rest of this section discusses the dimensions with large or significant differences.

For Dimensions 1 and 2 the averages were also noticeably different. These suggest tendencies for the non-natives to speak rather more than the natives and to have less overlapped speech, but do not obviously involve construction-skill differences.

For Dimension 3, although there was only a tiny difference in means, the non-natives exhibited narrower variation. This suggests fewer (or less prototypical) examples of swift turn takes and turn yields.

For Dimension 4 the average was slightly lower. Dimension 4-hi involved peaks in the interlocutor’s volume, speaking rate, and creakiness about two seconds apart, with an interleaved short contribution by the other speaker, which was usually a backchannel, short question, suggestion of word that the other was looking for, or laugh. This suggests that the non-natives less commonly produced small utterances precisely interleaved in the other’s turn.

For Dimension 6 the non-natives averaged slightly higher, indicating a tendency to pause to think more. The prosody of Dimension 6-hi was complex, including a pause surrounded by
two regions of high volume, fast speaking rate, wide pitch range, and creakiness. In Example 2, these were realized on those two and yeah, it’s hard

Example 2 (nn011@35.6)
B: the CSS was fun too, but the PHP
A: PHP, yeah, you have to combine them, those two; yeah, it’s hard

For Dimension 10 the non-natives averaged lower, suggesting a greater tendency to align or agree with the other person; this also was observed. In Example 3 the words squeeze you (showing empathy by acting out how B probably felt, by addressing an imagined dog) are fast, creaky and in wide pitch range, and they leads into a high-pitched laugh. (Interestingly this construction resembles the Upgraded Assessment Construction discussed above, although the timing is not the same.)

Example 3 (nn007@126.8)
B: (about his attempt to use a dog as a pillow) I’m sorry but you’re so fluffy
A: laughs, I just want to use it as a pillow, and squeeze you, laughs

While our focus has been on the top 16 dimensions, we ran the statistics down further, and noted significant differences for some others. Space permits discussion of only two.

For Dimension 18 the non-natives had fewer low values, suggesting fewer positive-to-negative perspective shifts. Examples of these included (my favorite class is) programming languages, because it’s the only hope I have (to get an A), with the last clause wry in tone, and the material’s really easy, so a lot of people, like stop paying attention to the class, and that’s what I did (and that’s why I failed it last time). Prosodically the 18-low construction involves a region of high pitch and high pitch range, followed directly by a region of low pitch.

For Dimension 21 the non-natives averaged lower. Dimension 21-lo was associated with filler production while recalling something from memory, where the filler was flat in pitch and initially creaky. 21-hi was associated with a rushed start to grab or hold a turn, with wide pitch range, and often followed by a reformulation. Fillers were indeed common in the non-native utterances, and aggressive turn starts rare.

In every case, the differences in behaviors suggested by the statistical analysis were confirmed by listening to the data and noting the common dialog activities, stances, and behaviors of the non-natives.

While the differences found are interesting, it is probably more significant that for most of the functions on our list there were only miniscule differences in means. Thus this method failed to suggest many expected construction-specific skill deficits. Indeed, even when differences were observed, there are obvious alternative explanations. The increased use of Dimensions 10-lo (alignment) and 6-hi (displaying empathy), can both be related to well-attested norms of Mexican culture (Condon, 1985). It seems likely that the non-natives were behaving as they thought appropriate, rather than trying to behave like natives but failing due to a prosodic skill deficit. Similarly, the reduced use of Dimension 4-lo (aggressive/rushed turn holding), could be explained as a choice not to use (or not to acquire) a behavior that can seem rude. The
reduced prevalence of 4-lo, 3-lo, and 3-hi (interpolated short contributions, and swift turn takes and yields) may reflect processing limitations: non-natives may be slow to comprehend and/or need more time to create fluent utterances (Wiberg, 2003).

In sum, this approach revealed effects of interaction style and adherence to cultural norms for conversation behavior (Tannen, 1989), but no real evidence for prosodic-skill deficits.

However lack of differences in the distributions does not mean a lack of difference in behavior. We realized this when an examination of the distributions of our Spanish data on the same dimensions revealed only minor differences. We speculate that this is because the space of possible prosodic variation is limited, and so languages tend to use the entire space, although for different purposes. Be that as it may, it is clear that examining distributions is not adequate for discovering all nonnative prosody differences.

7. Approach 2: Comparing Dimensions

Given the limitations of the previous method, we tried a second approach. Rather than directly measuring nonnatives’ behavior against the natives dimensions, this compares the patterns of non-native behavior to the native patterns. Thus, in this approach the first step is to characterize the prosodic behaviors of the non-natives in their own terms.

Figure 7 shows the concept. For the comparison step, the assumption is that, if the nonnatives’ behavior is similar to that of the natives in some respect, then the relevant pattern of native behavior will be well matched by some nonnative behavior pattern. Conversely, we assume that native patterns that lack a counterpart nonnative pattern will correspond to behaviors that the nonnatives have not mastered. To reduce the extent of “muddying” due to the behaviors of the native-speaker partners, for the PCA we ensured that the non-natives were always in the A track.

Since we use patterns based on PCA-derived dimensions, finding counterparts is easy. Each dimension is defined by its loadings on the raw features, so two dimensions are similar to the extent that their loadings are similar. We use the simplest operator for this, the cosine. Table 5 shows the cosines between the top five reference-native dimensions and the top five non-native dimensions.
Thus this method uses PCA to reduce the high-dimensional space of all prosodic features to a lower-dimensional space in which the patterns of the two speaker populations can be meaningfully compared. As this method is entirely form-based it could, in theory, be fooled if the non-natives use patterns that perfectly match native patterns in form, but are intended to convey different meanings. However each construction involves prosodic features for both speakers, so this would only be a problem if both speakers conspired in such perverse behavior; this was never observed.

Table 5 shows the result: the cosines between the top five reference dimensions and the top five non-native dimensions.

7.1 Major Differences

From Table 5 it is clear that top four dimensions do have counterparts, but for Reference Dimension 5 there is no strongly-similar non-native dimension. To identify other reference dimensions without a clear non-native counterpart, we computed Table 6, showing the cosines of the best matching dimensions from both the non-native data and the native-data comparison set. As might be expected, this second population of native speakers does not show exactly the same prosodic behavior as the reference population: column two is never 1.00. At the same time, also unsurprisingly, the natives are almost always closer to the reference than are the non-natives.

The table also shows that for many dimensions the non-native differences are minor; this is understandable given the speakers’ advanced level. However there are major differences for Dimensions 5 and 7.

We accordingly examined in detail Dimensions 5 and 7, first to delineate their function in natives’ dialogs, then to examine how the non-natives differed.

For Dimension 5, the low-side pattern involves about a half-second of increased volume, starting with high pitch and ending creaky. At times in the data when Dimension 5 was most strongly negative we frequently saw discourse markers, such as yeah, ah, ooh, and but, being used assertively. In general, when Dimension 5 is negative the speaker is showing involvement. The high-side pattern of Dimension 5 involves low pitch over several seconds, and within that a short utterance that is fast, creaky, and even more strongly low in pitch. This frequently occurring with words like and, um, like, and you know, for example when one speaker is musing about his future plans. In general, at times when Dimension 5 was high, the speaker had low involvement in the topic and/or the dialog itself.

The lack of a non-native dimension corresponding to Reference Dimension 5 suggests that the non-natives were not appropriately using prosody to indicate involvement and the lack thereof. To confirm this we examined their data in two ways. First, we listened to what was happening in each of the non-native dialogs at times when Reference Dimension 5 was strongly negative or strongly positive. On the negative side, we found that only some of the non-natives used this prosodic pattern to show involvement. One of the non-natives appeared not to use it at all; that is, there were no times where her speech was highly negative on this dimension. Another used this prosody frequently on question-initial so, making her questions sound incongruously aggressive. On the positive side, while some non-natives used this prosodic pattern sometimes for low involvement, they also used it in other contexts, for example in offering help, in marking disfluencies, and in greeting. As a second way to help understand what was going on, we examined the functions of the non-native dimensions which were (somewhat) similar to
this dimension, namely 4, 5, and 6. Their functions included the co-construction of utterances, floor holding, backchanneling, and marking the point of a story, but not involvement. Thus, there is good evidence that the method identified a real weakness.

For Dimension 7, the low-side pattern involves pitch strongly high across about 3 seconds, and in the middle of that a region with a fairly slow drop in volume, rate and creakiness over about 1.5 seconds. Native speakers frequently use this pattern to solicit empathy, and sometimes also when leaving something unsaid and inviting the listener to infer it. The high-side pattern of Dimension 7 involves strongly low pitch over about 3 seconds. When native speakers used this they were generally explaining something, usually something factual, such as a software project’s architecture, or how a study group had arranged to turn in a joint assignment.

To investigate the significance of there being no non-native dimension corresponding to Reference Dimension 7, we started by considering the negative side. There were many cases where the non-natives were using essentially this same pattern for essentially the same function: soliciting empathy, understanding, or an inference. Thus there was no apparent deficit on the negative side. On the positive side, however, we found no cases where non-natives used long regions of low pitch in the course of explaining things. It is not that they never explained technical things; rather they tended to do so in an interactive style, including lots of pitch variation, for example on interleaved questions to check that the listener was following. Some non-natives didn’t use the long low-pitch region pattern at all; others did, but used it not for explaining but when talking about something personal, such as family background, likes and dislikes, habits, or intentions.

### 7.2 Minor Differences

The method is effective not only for identifying gaps in learners’ skills, but also for detecting where the non-natives are using essentially the same constructions in the same ways, but with small differences. These can be inferred from differences in the loadings of corresponding dimensions.

For example, non-natives’ Dimension 1 is very similar in loadings to reference Dimension 1, and in the recordings they obviously serve the same function: positive when the left speaker has the floor, and negative when the right speaker does. However loadings are not identical, and in particular for some speaking rate features the loadings are higher for the natives. The difference is that the natives tend to speak faster when they clearly have the floor, but the non-natives have no such tendency. (To investigate whether this might be due to transfer from Spanish, we also ran PCA on the Spanish data: the corresponding dimension there indeed lacked a tendency for the person holding the floor to speak faster than average.)

Reference Dimension 2 and Non-native Dimension 2 also differ slightly in loadings, indicating that on the high side, during regions of overlapped talk by the two speakers, the natives tend to have a fast speaking rate, whereas the non-natives tend to have a higher pitch. For Reference Dimension 3 (turn hand-offs, Figure 1) and Non-native Dimension 3, the major difference is that the natives tend to speak faster at turn starts, but for non-natives this tendency is much weaker.

While these investigations suggest that the method is valid, in the sense that the differences that it uncovers are real ones, it appears not to be reliable for less-frequent constructions. Look-
ing again at Table 6, it is clear that the lower-ranked reference dimensions tend to align less well to the dimensions of the other data sets. This likely reflects a lack of robustness to extraneous sources of variability. This can be seen in the results for Reference Dimension 12. This pattern involves one speaker interleaving a short comment during a brief pause by the other, often showing alignment, appreciation, or empathy. While the non-native population appears to be doing this fairly successfully (a cosine of .79), the comparison native population appears to lack this pattern (highest cosine of .62). Looking at the data, we believe that this reflects not differences in prosodic competence but rather the fact that the comparison natives tended to talk more about technical topics than about personal ones, giving them fewer opportunities to be supportive.

8. Summary and Open Questions

We have presented methods able to find, from dialog data from a reference population and a non-native population, the ways which their prosodic behaviors most differ. Software for this workflow is available as open source (Ward, 2015).

Further, we applied these methods to recordings of Spanish-native speakers of English, identifying several respects in which their usage of prosodic constructions was different, including those relating to turn-taking and showing involvement.

It is interesting to speculate about how these prosodic differences may relate to perceived cultural differences. American businessmen often perceive Mexicans, it has been said, as being leisurely and disinclined to rush, and as tending to bring personal and emotional considerations into business discussions, rather than rationally sticking to facts (Condon, 1985). In the discussion of Dimension 1 we noted that the nonnatives do not tend to pick up the pace of speaking even when they have the floor, and the discussion of Dimension 7 implies that the nonnatives do not tend to consistently mark factual, explanatory information differently from personally-relevant information. Thus, while there may be real cultural differences, these cross-cultural perceptions may also reflect differences in prosodic behavior.

To evaluate our methods, since there exists no available set of judgments of dialog-prosody deficits, or even the knowledge needed to create such a set, we instead examined the validity of the differences that it found. The findings suggested that the methods are valid, but better evaluation is needed. While judgments of intended pragmatic are unavoidably subjective, at least independent judgments from more observers should be used. Among other things, this would enable eventual tuning of the feature set and similarity metric so that the resulting dimension orderings and difference ratings better reflect what is perceptually most salient and communicatively most important.

This leaves many open questions, including:

What are the details of each construction, and how do they work together in actual dialog? Here we relied heavily on automated methods, seeking a big-picture inventory and broad-brushstroke understanding. Like many other big-data methods this was efficient (Swanson and Charniak, 2014), but has its limits. Further examination using more sensitive methods could better tie these construction-based descriptions to those developed within other theoretical frameworks. The resulting detailed understanding of dialog-prosody would be invaluable for many purposes, including second-language teaching.
Which non-native differences in prosody matter? On the one hand, differences may “make the speaker sound strange, typical of their origin, boring or annoying . . . [but] . . . not cause much of an actual breakdown in communication” (Wells, 2014). On the other, such differences may affect perceptions and dialog outcomes (Tannen, 2005; Curhan and Pentland, 2007). Identifying which differences matter will require both broader consideration of interpersonal and social factors, and also more work on the nuts and bolts of dialog prosody.

How can we help non-natives master the prosodic constructions of a new language? There are many techniques for teaching prosody, but learning and teaching interaction patterns in dialog involves special challenges (Betz and Huth, 2014).

Finally, while this paper examined only general patterns of behavior across a population of non-native speakers, we would like to explore using this method to pinpoint individual speakers’ deficits. Among other challenges, this would require investigating how to obtain reliable results with less data. If this can be done, these methods should support not only discovery of differences but also assessment of learners’ dialog-prosody proficiency and diagnosis of deficiencies.

Acknowledgments

We thank the participants who let us record their conversations, and Richard Ogden, David Novick, David DeVault, Juergen Trouvain, Francisco Torreira, and Shizuka Nakamura for discussion. This work was supported in part by the US National Science Foundation by Research Experience for Undergraduates supplements to IIS 191-4868 and IIS 144-9093, and in part by the Fulbright Program.

References


Anais G. Rivera and Nigel Ward. Prosodic cues that lead to back-channel feedback in Northern Mexican Spanish. HDLS-7 Conference, High Desert Linguistics Society, University of New Mexico, 2006.


<table>
<thead>
<tr>
<th></th>
<th></th>
<th>lo:</th>
<th>hi:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17%</td>
<td>Speaker A speaking, Speaker B silent</td>
<td>Speaker B speaking, Speaker A silent</td>
<td>§3.4, §6, §7</td>
</tr>
<tr>
<td>2</td>
<td>8%</td>
<td>both speakers silent</td>
<td>both speakers talking together or laughing together</td>
<td>§6, §6, §7</td>
</tr>
<tr>
<td>3</td>
<td>4%</td>
<td>B yields the turn and A takes the turn</td>
<td>A yields the turn and B takes the turn</td>
<td>§4, §6, §7</td>
</tr>
<tr>
<td>4</td>
<td>3%</td>
<td>B makes a small contribution during A’s turn</td>
<td>A makes a small contribution during B’s turn</td>
<td>§6</td>
</tr>
<tr>
<td>5</td>
<td>3%</td>
<td>high involvement</td>
<td>low involvement</td>
<td>§7</td>
</tr>
<tr>
<td>6</td>
<td>3%</td>
<td>pivot point of a rhetorical structure, etc.</td>
<td>pausing while thinking how to continue</td>
<td>§6</td>
</tr>
<tr>
<td>7</td>
<td>3%</td>
<td>bidding for empathy, inviting an inference</td>
<td>giving factual information, explaining something or someone’s actions</td>
<td>§7</td>
</tr>
<tr>
<td>8</td>
<td>2%</td>
<td>A confident, speaking with authority or based on personal experience</td>
<td>B confident, speaking with authority or based on personal experience</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2%</td>
<td>A disfluent, hesitant, or silent; B silent or fluent</td>
<td>B disfluent, hesitant, or silent; A silent or fluent</td>
<td>§6, §6</td>
</tr>
<tr>
<td>10</td>
<td>2%</td>
<td>speakers aligning</td>
<td>speakers disagreeing or diverging</td>
<td>§4</td>
</tr>
<tr>
<td>11</td>
<td>2%</td>
<td>B yields floor to A</td>
<td>A yields floor to B</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1%</td>
<td>B interpolates a short comment</td>
<td>A interpolates a short comment</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1%</td>
<td>lack of new information</td>
<td>knowledge asymmetry between speakers</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1%</td>
<td>personal-situation comments</td>
<td>complaints about third parties</td>
<td>§6</td>
</tr>
<tr>
<td>15</td>
<td>1%</td>
<td>being positive about one’s own prospects or a past experience</td>
<td>negative feeling about something/someone distant</td>
<td>§6</td>
</tr>
<tr>
<td>16</td>
<td>1%</td>
<td>displeasure, annoyance</td>
<td>amusement, positive evaluation of something/someone</td>
<td>§4</td>
</tr>
<tr>
<td>18</td>
<td>1%</td>
<td>A reveals downside or B reveals silver lining</td>
<td>B reveals downside or A reveals silver lining</td>
<td>§6</td>
</tr>
<tr>
<td>21</td>
<td>1%</td>
<td>memory recall</td>
<td>rushed turn grab or hold</td>
<td>§6</td>
</tr>
</tbody>
</table>

Table 2: The top sixteen prosodic dimensions in the reference corpus, plus two more. The second field is the amount of variance explained by the dimension. The third field summarizes our interpretations of the dimension when negatively or positively present, that is, the “lo-side” and “hi-side” constructions, as discussed below. The fourth field indexes further discussion.
Table 3: Summary of the Data Sets

<table>
<thead>
<tr>
<th>use</th>
<th>language</th>
<th>speakers</th>
<th>amount used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Speech</td>
<td>building the model</td>
<td>English</td>
<td>English-native</td>
</tr>
<tr>
<td>Switchboard</td>
<td>testing approach 1</td>
<td>English</td>
<td>English-native</td>
</tr>
<tr>
<td>UTEP nonnative English</td>
<td>finding skill deficits</td>
<td>English</td>
<td>Spanish-native</td>
</tr>
<tr>
<td>UTEP monolingual English</td>
<td>calibrating differences</td>
<td>English</td>
<td>English-native</td>
</tr>
<tr>
<td>UTEP Spanish</td>
<td>exploring L1 influences</td>
<td>Spanish</td>
<td>Spanish-native</td>
</tr>
</tbody>
</table>

Table 4: Statistics on dimension use in three data collections, relative to the reference set. For the reference set, not shown, the means are all 0 and the standard deviations all 1, due to normalization. Bolding marks the largest differences between Switchboard and the reference set, and the largest differences between the non-natives and the natives. Asterisks indicate significant native/non-native differences.

<table>
<thead>
<tr>
<th>dimension</th>
<th>Switchboard mean</th>
<th>std. dev.</th>
<th>natives mean</th>
<th>std. dev.</th>
<th>non-natives mean</th>
<th>std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.00</td>
<td>1.22</td>
<td>-0.00</td>
<td>1.05</td>
<td>-0.11</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>-0.31</td>
<td>0.53</td>
<td>-0.33</td>
<td>0.76</td>
<td>*-0.45</td>
<td>0.75</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>0.91</td>
<td>0.00</td>
<td>0.91</td>
<td>0.01</td>
<td>0.86</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.87</td>
<td>-0.00</td>
<td>0.85</td>
<td>*-0.05</td>
<td>0.83</td>
</tr>
<tr>
<td>5</td>
<td>-0.12</td>
<td>0.74</td>
<td>-0.01</td>
<td>0.94</td>
<td>0.01</td>
<td>0.91</td>
</tr>
<tr>
<td>6</td>
<td>-0.17</td>
<td>0.65</td>
<td>0.10</td>
<td>0.86</td>
<td>0.14</td>
<td>0.84</td>
</tr>
<tr>
<td>7</td>
<td>-0.06</td>
<td>0.86</td>
<td>-0.06</td>
<td>0.97</td>
<td>-0.03</td>
<td>0.97</td>
</tr>
<tr>
<td>8</td>
<td>-0.00</td>
<td>0.81</td>
<td>0.00</td>
<td>0.86</td>
<td>0.03</td>
<td>0.86</td>
</tr>
<tr>
<td>9</td>
<td>0.00</td>
<td>0.85</td>
<td>0.00</td>
<td>0.82</td>
<td>-0.00</td>
<td>0.81</td>
</tr>
<tr>
<td>10</td>
<td>-0.43</td>
<td>0.97</td>
<td>-0.10</td>
<td>1.01</td>
<td>*-0.21</td>
<td>0.99</td>
</tr>
<tr>
<td>11</td>
<td>-0.00</td>
<td>0.94</td>
<td>-0.00</td>
<td>0.91</td>
<td>-0.01</td>
<td>0.93</td>
</tr>
<tr>
<td>12</td>
<td>-0.00</td>
<td>0.92</td>
<td>-0.00</td>
<td>0.89</td>
<td>-0.00</td>
<td>0.89</td>
</tr>
<tr>
<td>13</td>
<td>0.11</td>
<td>0.71</td>
<td>0.03</td>
<td>0.91</td>
<td>0.03</td>
<td>0.90</td>
</tr>
<tr>
<td>14</td>
<td>-0.33</td>
<td>0.79</td>
<td>-0.07</td>
<td>0.94</td>
<td>-0.09</td>
<td>0.95</td>
</tr>
<tr>
<td>15</td>
<td>0.37</td>
<td>0.87</td>
<td>0.07</td>
<td>0.97</td>
<td>0.07</td>
<td>0.98</td>
</tr>
<tr>
<td>16</td>
<td>0.00</td>
<td>0.88</td>
<td>-0.00</td>
<td>0.92</td>
<td>-0.02</td>
<td>0.92</td>
</tr>
<tr>
<td>17</td>
<td>-0.00</td>
<td>1.02</td>
<td>-0.00</td>
<td>0.97</td>
<td>*0.11</td>
<td>0.91</td>
</tr>
<tr>
<td>18</td>
<td>0.00</td>
<td>0.95</td>
<td>0.00</td>
<td>0.88</td>
<td>*-0.13</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 5: Similarities between reference dimensions and non-native dimensions. In each row the highest value is bolded.
<table>
<thead>
<tr>
<th>reference dimension</th>
<th>cosine of the most strongly-related dimension native</th>
<th>cosine of the most strongly-related dimension non-native</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.99</td>
<td>.97</td>
<td>.98</td>
</tr>
<tr>
<td>2</td>
<td>.90</td>
<td>.90</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>.92</td>
<td>.95</td>
<td>1.03</td>
</tr>
<tr>
<td>4</td>
<td>.92</td>
<td>.80</td>
<td>.87</td>
</tr>
<tr>
<td>5</td>
<td>.80</td>
<td>.61</td>
<td>.76</td>
</tr>
<tr>
<td>6</td>
<td>.64</td>
<td>.63</td>
<td>.98</td>
</tr>
<tr>
<td>7</td>
<td>.83</td>
<td>.57</td>
<td>.69</td>
</tr>
<tr>
<td>8</td>
<td>.73</td>
<td>.63</td>
<td>.86</td>
</tr>
<tr>
<td>9</td>
<td>.74</td>
<td>.69</td>
<td>.93</td>
</tr>
<tr>
<td>10</td>
<td>.67</td>
<td>.64</td>
<td>.96</td>
</tr>
<tr>
<td>11</td>
<td>.87</td>
<td>.83</td>
<td>.95</td>
</tr>
<tr>
<td>12</td>
<td>.62</td>
<td>.79</td>
<td>1.27</td>
</tr>
<tr>
<td>13</td>
<td>.74</td>
<td>.72</td>
<td>.95</td>
</tr>
<tr>
<td>14</td>
<td>.75</td>
<td>.64</td>
<td>.85</td>
</tr>
<tr>
<td>15</td>
<td>.74</td>
<td>.68</td>
<td>.92</td>
</tr>
<tr>
<td>16</td>
<td>.60</td>
<td>.56</td>
<td>.93</td>
</tr>
</tbody>
</table>

Table 6: For each of the reference dimensions, the cosine of the best-matching dimension found for the other data sets.