Non-Native Differences in Prosodic-Construction Use

Nigel G. Ward NIGELWARD@ACM.ORG
Department of Computer Science
University of Texas at El Paso and Kyoto University
El Paso, TX, 79968, USA

Paola Gallardo PGALLARDO@MINERS.UTEP.EDU
Department of Computer Science
University of Texas at El Paso
El Paso, TX, 79968, USA

Editor: Dr. Amanda Stent

Abstract

Many language learners never acquire truly native-sounding prosody, and often are weak on the dialog-related uses of prosody. Previous work has suggested this may involve deficits with specific prosodic constructions, but this has not been systematically investigated. We developed semi-automatic analysis methods able to identify and characterize such differences. Starting with two sets of dialog data, one of native speakers and one of non-natives, we applied Principal Components Analysis, and then identified differences in distributions and in the constructions themselves. Applied to recordings of six advanced-level native-Spanish learners conversing in English, these methods revealed differences in their uses of speaking rate and pitch in turn-taking, and infrequent and variant use of the English prosodic constructions for showing involvement and for explaining.

Keywords: conversation, dialog, interaction, pragmatics, language learning, non-native prosody, comparison, second language, L2, American English, Mexican-Spanish speakers, L1 transfer, turn-taking, principal components analysis, semi-automatic methods

1. Introduction

Non-native speakers often have saliently non-native prosody, even when their other language skills are good (Zimmerer et al., 2014; Mennen, 2015). 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 exploration of these issues.

The contributions are 1) semi-automated methods for discovering prosodic differences from data, 2) an inventory of some important prosodic constructions of English, 3) descriptions of the prosodic forms and pragmatic functions of some of these constructions, and 4) the identification of three ways in which the prosody of native-Spanish learners of English differs from that of native English speakers.

In this article our interest is in "dialog prosody," by which we mean uses of prosody that help coordinate the interaction and convey attitude, pragmatics, and related functions. We are not here interested in the more-commonly studied aspects of prosody — prosody as it relates to words, syntax, semantics, emotion, and paralinguistics — although of course all uses of prosody are complexly interrelated (Couper-Kuhlen and Selting, 1996b; Hirschberg, 2002; Szczepek Reed, 2012; Wichmann, 2014).

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 non-native corpora existing, we collected 90 minutes of data from six advanced native-Spanish speakers in English conversations with native English speakers (Section 5). Simple statistical measures revealed some differences in prosodic construction use (Section 6). Model-based comparison revealed others: for this we first identified the non-native prosodic constructions, and then compared these to the native speakers' constructions (Section 7). In Section 8 we summarize and note questions for future research.

2. Previous Methods for Characterizing Non-Native Prosodic Differences

It is not easy to accurately characterize how non-native prosody differs from native-speaker prosody, especially for the dialog-related uses. This section overviews four commonly used approaches and 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), Spanish turn keeping and information seeking (Aronsson and Fant, 2014), and backchanneling in German and Vietnamese (Ha et al., 2016). Such analyses rely on the existence of a clear intended function, and thus they cannot give us the whole story. One reason is that 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. Another reason is that such studies can only discover differences for previously-identified functions, and there is no guarantee that the pragmatic functions identified to date are exhaustive, or even cover those most important in actual conversation.

A second approach starts with a syntactic form or sentence type and looks at differences between learners' and native speakers' 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, prosody is often determined more by the pragmatic function than the syntactic form, especially in dialog (Lai, 2012; Hedberg et al., 2014), so 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 the 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, and these are always limited. For example, the most popular current models of prosodic forms are based on monolog data, and they mostly handle only pitch (intonation), leaving out speaking rate, timing, and intensity, although these aspects are also 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 showed that non-native speakers use much less pitch variation (Zimmerer et al., 2014). This method can exploit large amounts of data and is entirely objective. It is also robust: while in any given utterance a non-native speaker may have good reason to use compressed pitch range — for example when losing interest in a topic and preparing to close it out — consistently limited pitch variation across a corpus is good evidence for a real difference. 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 a wider pitch range would have been appropriate.

These methods have provided insights regarding many specific differences in prosody (Mennen, 2015). Nevertheless each has its weaknesses, and thus we wish to explore a new way to investigate non-native prosody. Our approach, like the latter two above, starts with forms, rather than with functions, and is corpus-based. Because we are centrally interested in what people actually do in conversation, we wish to discover from the data itself what functions are expressed with prosody, and how non-native speakers differ.

3. Prosodic Constructions and their Automatic Discovery

Describing prosodic behavior is difficult, and there is currently no consensus on how to represent prosodic knowledge. Rather there are many different approaches, with different assumptions, methods, and descriptive vocabularies. Several good surveys exist (Cutler and Ladd, 1983; Couper-Kuhlen, 1986; Ladd, 1996; Wells, 2006; Szczepek Reed, 2006; van Santen et al., 2008; Arvaniti, 2011; Xu, 2011; Prieto, 2015). 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, for two reasons: it supports semi-automated analysis, and it directly represents the prosodic forms associated with pragmatic functions. This section explains the notion of prosodic construction and how we discover the constructions of a language from a collection of dialog recordings.

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 intensity, pitch height, and pitch range, 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.

time	Speaker A	Speaker B
–3000 to –300ms	speaking	quiet
–2500 to –1300	speaking louder	quiet
–800 to 0	tapering in loudness to silence	
–300 to +300	quiet or silent	loud and fast
+300 to +600	resuming speaking	slowing and fading out
+600 to +1300	speaking loudly	quiet
+1300 to +3200	speaking	quiet

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.

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 include other prosodic features, such as intensity, rate, and timing: thus they are multistream models. 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. This seems descriptively necessary, and is an essential part of many modern models of prosody (van Santen et al., 2004; Chen et al., 2004; Xu, 2005). 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 its 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 and measurements of construction differences.

3.2 Prosodic Construction Discovery by Principal Components Analysis

In order to examine non-native 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 over 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, including some based on clustering, Functional Data Analysis (FDA), and PCA (Itahashi and Tanaka, 1993; Chen et al., 2005; Gubian et al., 2011; Parrell et al., 2013; 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.

In this particular study, the datapoints are taken every 10 milliseconds throughout the conversations. This means that, rather than considering prosody only at turn-ends, or only when computed over utterances, as is done in many approaches, the method considers prosody as it appears everywhere in the conversations. (Indeed, datapoints are taken even during silent regions. This makes sense, because silence is also a dialog phenomenon, and can be part of larger patterns of behavior, as will be seen.)

3.3 Prosodic Features Used

This subsection documents the prosodic features used as input to PCA. Since our interest is in dialog prosody, we wanted to use features relevant to the prosodic forms involved in expressing dialog-related functions. While there is no sharp distinction between the prosodic forms involved in in different functions — for example, the same feature can convey either lexical identity or pragmatic function, depending on the language or the speaker — there are tendencies. In particular, it seems that prosodic features which are anchored to or aligned with other linguistic units – syllables, words, sentences — tend to relate more to lexical and syntactic functions. We therefore use unaligned features. This does not mean that other prosodic functions will be entirely excluded from our models, but it does reduce their effects.

The features were computed at every timepoint, without relying on any segmentation of the input. While prosodic analysis is often done subsequent to a segmentation of the input, for example into turns, here we do without such preprocessing. One reason is that some prosodic patterns are not turn-aligned, so if we restricted attention to turn-aligned features, the analysis would likely not find such patterns. Another reason is that, although the notion of "turn" seems straightforward, in spontaneous conversations turns are difficult to identify reliably, even by humans annotators following strict guidelines, so by avoiding this step, we simplify the process.

Since we need features that can support the discovery of temporal patterns, for each datapoint we used a number of features at different offsets to broadly represent the local prosodic context. For example, in addition to the intensity over the past 50 milliseconds, we also used the intensity 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. Following previous work, we chose windows of various sizes so as to give greater temporal resolution near the time of interest, that is the timepoint at the center of all the features.

Since our interest is in prosody, not just intonation, we included not only pitch features but also features for speaking rate, intensity, and creaky voice. While there are many more features that could be included, this set was designed to capture most of the prosodic information that has been found most useful for many tasks (Schuller, 2011; Shriberg and Stolcke, 2004; Ward et al., 2011). We followed previous work in having more windows and finer resolution for the features that usually are most informative, notably intensity.

In total we used 176 features, as listed in Figure 1. These were computed using our open-source toolkit (Ward, 2015). This includes built-in normalizations to make the features fairly speaker-independent. For loudness we used log energy normalized per track to correct for different recording conditions and different speakers. For the pitch-height and pitch-range features we used percentiles in the distribution of pitch seen for that track, thus again normalizing for speaker. For speaking rate, we used a simple frame-by-frame energy-difference measure. To avoid the problems associated with interpolating pitch over nonvoiced regions, we used features representing the strength of evidence for the pitch being low (respectively, high, narrow, wide) over a region, using evidence computed only over valid pitch points. For example, our narrow-pitch feature counts the number of pairs of pitch points within a window between which the pitch varies less than 2%. This feature, like the others, was designed to be robust: to roughly match perceptions over a great variety of voices, dialog activities and noise levels. None are entirely reliable, and in particular the speaking-rate proxy, although intended to detect fast speech versus lengthening, also responds to precise articulation (enunciation) versus phonetic reduction, and to creaky versus modal voice. This feature set was designed and refined based largely on experience with various prediction tasks (Ward and Vega, 2012; Ward et al., 2011); experience also shows that minor changes to the feature windows or feature implementations have little effect on the dimensions that PCA finds, doubtless due to their overall robustness and to the size of the dataset used.

amplitude (16 per speaker)	low pitch, high pitch, creakiness (14 each, per speaker)	narrow pitch, wide pitch (10 each, per speaker)	speaking rate (10 per speaker)
-3200 – -1600			
-1600 – -800	-1600 – -800	-1600 – -800	-1600 – -800
-800 – -400	-800 – -400	-800 – -400	-800 – -400
-400 – -300	-400 – -300	-400 – -300	-400 – -200
-300 – -200	-300 – -200	-300 – -200	
-200 – -100	-200 – -100	-200 – 0	-200 – -100
-100 – -50	-100 – -50		-100 – 0
-50 – 0	-50 – 0		
0 – 50	0 – 50		
50 – 100	50 – 100		0 – 100
100 – 200	100 – 200	0 – 200	100 – 200
200 – 300	200 – 300	200 – 300	
300 – 400	300 – 400	300 – 400	200 – 400
400 – 800	400 – 800	400 – 800	400 – 800
800 – 1600	800 – 1600	800 – 1600	800 – 1600
1600 – 3200			

Figure 1: The prosodic feature inventory. Start and end times for each window in milliseconds offset from the point of interest. These features are computed for these windows for both left and right speakers, giving 176 in total.

3.4 From Dimensions to Constructions

The workflow is summarized in Figure 2. For each timepoint in the corpus the prosodic features are computed. PCA digests all this data and outputs 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 configuration it will be positive. Every dimension codes for two patterns in this way: one when it is present positively, and one when it is present negatively.

Dimensions usually have, moreover, temporal variation in the loadings. That is, a certain feature, like high pitch, may be indicative of a pattern being present when it occurs at one time, but not at another. 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. This particular pattern of loadings is easy to understand: it corresponds to the well-known prosodic phenomenon of declination. The fact that a simple mathematical operation, PCA, can find such a pattern, even though we were not looking for it, illustrates its power. Because all patterns observed so far involve extensive temporal variation in loadings, and thus represent temporal configurations of features, it is appropriate to call them constructions, as we will henceforth.

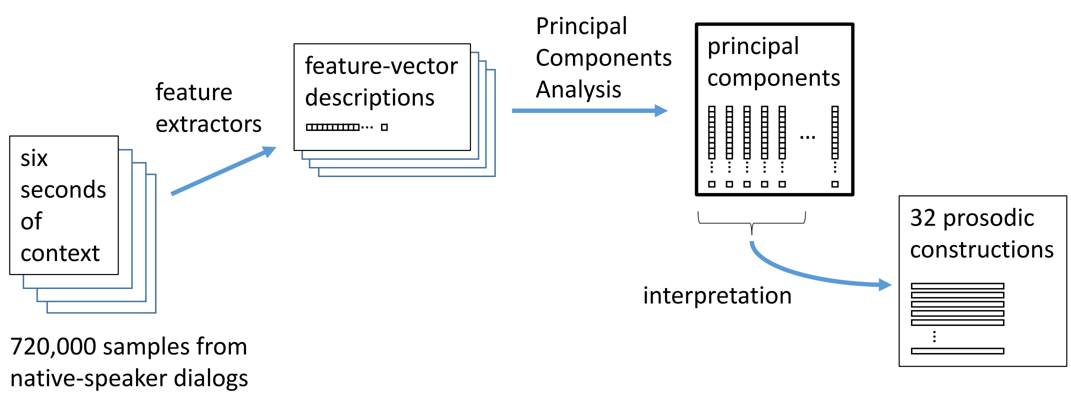

Figure 2: Principal Components Analysis workflow.

3.5 Inferring Construction Meanings

We are interested not only in prosodic patterns, but also in what they mean. While some prosodic patterns, like declination, may just be facts of language (or of a specific language), many have meaning or serve a function. To fully understand how learners' prosody differs from that of native speakers, we first need to understand these meanings. This is, however not easy: identifying meanings for prosodic constructions is rife with methodological pitfalls (Arvaniti, 2011; Prieto, 2015). Therefore our identifications of meanings for constructions must be regarded as tentative. This section starts with illustrations, and then describes the interpretation process systematically.

As mentioned above, one outcome of PCA is a prosodic description of each construction: its weightings for 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. This information being overwhelming with 176 features, it is convenient to use visualizations. For example, Figure 3 shows the loadings for Dimension 3, showing 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 intensity, speaking rate, and creakiness, followed by a further increases on the latter two and a simultaneous drop in pitch. Table 2 gives a simplified summary of this construction. For reasons of space, this paper only discusses aspects of the loadings that are relevant to non-native speakers' differences, but all are available at `http://www.cs.utep.edu/nigel/l2english/`, both numerically and as visualizations.

time	Speaker A	Speaker B
–2000 ∼ –1600 ms	loud, fast, creaky	
–1600 ∼ –800 ms	louder, faster, creakier	
–800 ∼ 400 ms	pitch drops, quieter	
–400 ∼ 0 ms	quieter falling to silent	
0 ∼ 400 ms		silent or a quiet, tentative start
400 ∼ 800 ms		loud, creaky, fast, high in pitch
800 ∼ 1200 ms		feature values revert to typical

Table 2: Major components of a prototypical rendition of the basic turn hand-off construction of English. Times are in milliseconds relative to the point halfway between the original speaker's end and the new speaker's start.

Of course, no actual instance of turn yield will exactly match this pattern, due to the simultaneous presence of other, superimposed, constructions. However there are cases that match quite well: an example is seen in Figure 4, transcribed as Example 1. (Audio for the examples also is available at the above URL.)

Example 1 (soc008@165.1s)

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?

We note in passing that most of the properties of the turn hand-off construction, as found here, are also present in other descriptions. For example Gravano and Hirschberg (2011) found similar turn-yield tendencies involving rate, intensity, non-modal voice, and final pitch. It is also interesting to note that in Dimension 3 the loadings of features for the two speakers are nearly symmetric: past-future mirror images across the point of interest (0 milliseconds). We

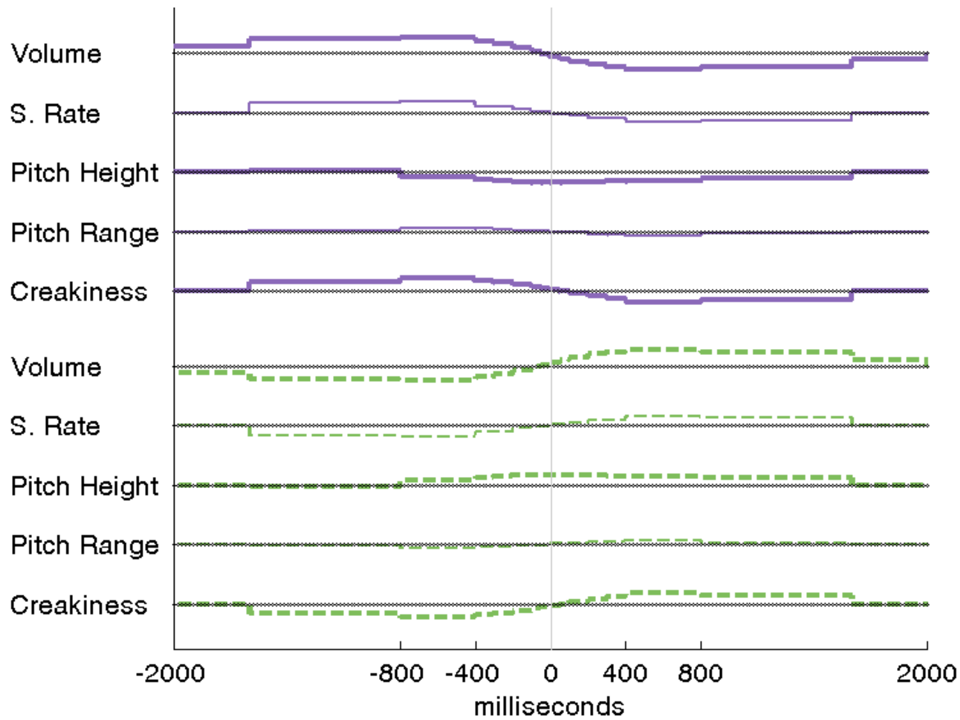

Figure 3: Loadings of Dimension 3. Purple solid lines are for the A speaker; green dashed lines for B. Time is in milliseconds. The dotted lines are zeros, with points above them indicating positively loaded features and points below negative. The "pitch height" line shows the difference between the 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 the figure suggests. While the intensity 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.

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

Dimension 3 was thus easy to understand, but this was not true for all dimensions. For most dimensions we couldn't infer the meaning from the loadings alone, so we relied more on examination of places in the corpus where a construction was strongly present. Specifically, we examined ten to twenty exemplars, places where the value on a given dimension was highest or lowest. For each of these we noted aspects of the context and dialog activity, and the pragmatic functions that were being expressed. These we inferred primarily from information in the dialog itself, including the words being said and the behaviors of both participants in the immediate context, a standard Conversation-Analysis technique (Sidnell, 2011). We also occa-

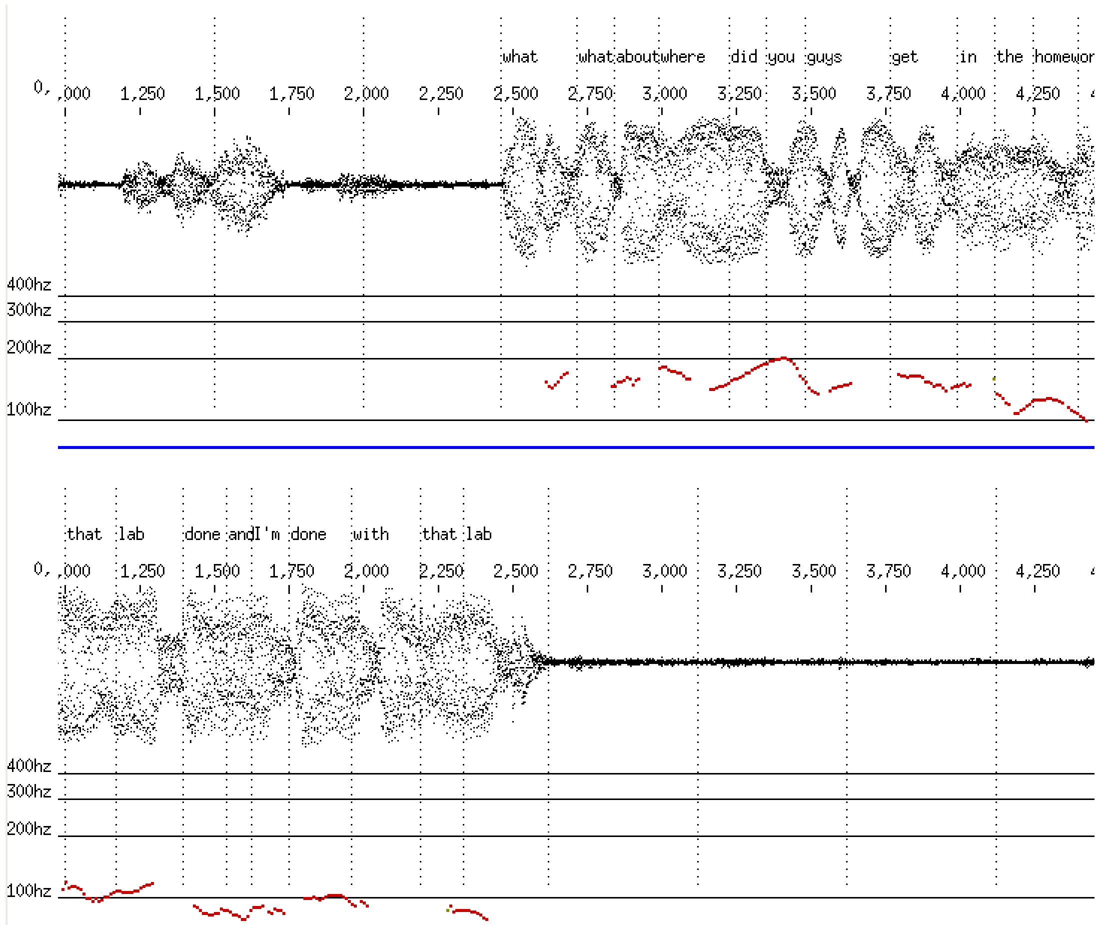

Figure 4: A swift turn exchange, high on Dimension 3. Pitch is shown at the bottom of each track.

sionally engaged our own intuitions about what was being conveyed by the observed prosodic form, sometimes by considering the contrast to other forms that the speaker might have used. We then used qualitative-inductive methods to find commonalities among the noted functions and meanings. For some dimensions the commonalities were obvious from just a few examples; for others the commonality did not become clear until we had examined many. In every case, after forming an initial hypothesized meaning, we examined more examples, either finding confirmation or, less often, discovering that we needed to refine or change it.

This interpretation process generally went smoothly. However, some of the exemplars were hard to relate to a general hypothesized meaning. One reason 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 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. It is also likely that our working assumption, that the meanings contributed by the different constructions are compositional, is not entirely accurate. A second thing that complicated interpretation was individual differences in behavior and uses. For example, for some of the high-side exemplars for Dimension 10, it seemed that the speaker was being provocative, not just simply disagreeing or diverging, the general functions. Quite often, the constructions appeared polysemous. As another example, for some of the low-side Dimension 10 exemplars, one speaker was producing a short check questions, or saying something they expected the other to agree with, and the other usually did. In the table we generalize over these and related meanings and refer to the function as "agreeing or aligning." Other analysts could make other choices for such summary phrases. A third complication for the analysis was the presence of creative and deliberate uses of prosody, including for non-literal meanings and for reported speech (Rao, 2013b; Estelles-Arguedas, 2015), many of which did not follow the general tendencies.

In a sufficiently advanced model, that properly accounted for all these factors, perhaps the prosody-meaning mappings could be seen to be functioning as exceptionless rules. As it is, our policy was to ascribe meanings to forms if those meanings were clearly present in most of the exemplars. Despite this imprecision, we see that PCA was effective in identifying factors (dimensions, patterns) that not only explain the observations, but also are meaningful. Of course, the demonstrated ability of PCA to do this for various types of data is why we chose it, so its effectiveness also for prosody is not a surprise (Ward and Vega, 2012; Ward, 2014; Ward et al., 2016).

4. Some Prosodic Constructions of English

To quantify the prosodic behavior patterns of the non-native speakers, we need to compare them to some standard. The language norm that our non-native speakers were most familiar with is General American English, especially as spoken by young people in the Southwest, so 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 university students studying computer science, although this was 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 the methods described above.

Table 3 summarizes the results. The second column shows the percentage of variance accounted for by each dimension. 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 number 16 has no special significance; we chose to stop at 16 due simply to limitations of time and space.) Our interpretations are summarized in the third column.

Ideally we would like a full and final listing of the prosodic constructions of English before going on to examine non-native speakers' 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 with different feature sets the resulting dimensions vary slightly (Ward and Vega, 2012; Ward, 2014). The descriptions are only suggestive, due to reasons of space, although each construction really deserves a paper in itself, to treat its form and function in detail, and to relate

1	17%	lo:	Speaker A speaking, Speaker B silent	§3.4, §6, §7.2, §8
		hi:	Speaker B speaking, Speaker A silent	§3.4, §6, §7.2, §8
2	8%	lo:	both speakers silent	
		hi:	both speakers talking together or laughing together	§3.5, §6, §7.2
3	4%	lo:	B yields the turn and A takes the turn	§3.5, §6, §7.2
		hi:	A yields the turn and B takes the turn	§4, §6, §7.2
4	3%	lo:	B makes a small contribution during A's turn	§6
		hi:	A makes a small contribution during B's turn	§6
5	3%	lo:	high involvement	§7.1, §8
		hi:	low involvement	§7.1, §8
6	3%	lo:	pivot point of a rhetorical structure, etc.	
		hi:	pausing while thinking how to continue	§6
7	3%	lo:	bidding for empathy, inviting an inference	§7.1, §8
		hi:	giving factual information, explaining something or some actions	§7.1, §8
8	2%	lo:	A confident, speaking with authority or based on personal experience	
		hi:	B confident, speaking with authority or based on personal experience	
9	2%	lo:	A disfluent, hesitant, or silent; B silent or fluent	
		hi:	B disfluent, hesitant, or silent; A silent or fluent	
10	2%	lo:	speakers agreeing or aligning	§3.5, §6, §8
		hi:	speakers disagreeing or diverging	§4
11	2%	lo:	B yields floor to A	
		hi:	A yields floor to B	
12	1%	lo:	B interpolates a short comment	§7.2
		hi:	A interpolates a short comment	§7.2
13	1%	lo:	lack of new information	
		hi:	knowledge asymmetry between speakers	
14	1%	lo:	personal-situation comments	
		hi:	complaints about third parties	
15	1%	lo:	being positive about one's own prospects or a past experience	
		hi:	negative feeling about something/someone distant	
16	1%	lo:	displeasure, annoyance	§3.5
		hi:	amusement, positive evaluation of something/someone	
18	1%	lo:	A reveals downside or B reveals silver lining	§6
		hi:	B reveals downside or A reveals silver lining	§6
21	1%	lo:	memory recall	§6
		hi:	rushed turn grab or hold	§6

Table 3: 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. The fourth field indexes further discussion.

it to alternate possible descriptions. We give detail, below, only for constructions that turn out to be used differently by non-native speakers. We also note that our method did not identify exclusively dialog-related aspects of prosody, nor, certainly, all 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 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, 1996a; Sidnell, 2011; Clark, 1996; Szczepek 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 of some generality.

5. Non-Native Dialog Data

For this study we chose to work with advanced non-native speakers, 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 (Van Engen et al., 2010). This section summarizes some of the important properties of our data sets; the details appear elsewhere (Ward and Gallardo, 2015).

We chose learners whose native language was Spanish, based on the ease of recruiting them. The segmental, lexical, and syntactic aspects of Spanish prosody are known to differ from English, as are 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 Colantoni, 2014). Accordingly it seemed likely that there would be differences in dialog prosody also.

We recruited among friends and acquaintances in our department; as a result, all participants had completed at least one semester of college in the United States. Participants gave informed consent, and we compensated them with $15 for participating. Each non-native speaker was recorded in dialog with an English-monolingual native-speaker partner. Later, after listening to the recordings, we decided to exclude four speakers who seemed to have almost-native conversation skills. Thus we obtained 9 conversations, including 6 different non-native speakers. These speakers all had strong vocabulary and good fluency, but all were noticeably non-native in pronunciation. All had grown up in Northern Mexico.

When recording we did not ask the speakers to do anything more specific than talk to each other. Their conversations were spontaneous and varied 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 prosody in monolog or scripted dialog is, in essence, “merely” a question of choosing the appropriate form 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, multiple constructions must often be simultaneously realized, and all this must be done under the time pressure of choosing words and listening to and coordinating with the dialog partner.

As noted above, we selected the non-native speakers for the corpus based on our perceptions of awkwardness with English, without explicit consideration of prosodic behaviors. However we did a post-hoc examination to see whether their prosody also appeared non-native. Casual listening showed that it was, 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.

We must here note two potential issues with this data. One is that, since each pair of speakers includes a native speaker, and since each pattern involves behavior by both speakers, it is possible that some observed differences could be due to the native speakers behaving differently when interacting with non-native speakers, rather than to differences in the behavior of the non-native speakers 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 speaker. Thus our method may understate the non-native differences.

For comparisons we used three other data sets. Two we recorded ourselves: 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. We also used native speakers from the well-known Switchboard corpus (Godfrey et al., 1992), specifically, 7 randomly-selected dialogs (14 speakers, 35 minutes total). Although these are also dyadic conversations in American English, in these conversations 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 conversations rapidly moved on to other topics. Table 4 summarizes the five data sets used.

	use	language	speakers	used
Social Speech	building the model, reference	English	English-native	60 min.
Switchboard	exploring distributions	English	English-native	35 min.
UTEP non-native English	finding skill deficits	English	Spanish-native	90 min.
UTEP native English	comparing difference magnitudes	English	English-native	70 min.
UTEP Spanish	exploring L1 influences	Spanish	Spanish-native	90 min.

Table 4: Summary of the data sets

6. Distribution Differences

We expected that the dialog-prosody deficits of non-native speakers could be associated with specific constructions. Logically, following Mennen's (2015) categories, these deficits could be of four kinds: not knowing a construction, using it too often or not enough, using it for the wrong functions, or not producing it accurately.

In our first approach we looked for differences of the first two kinds by comparing distributions on the various dimensions. If non-native speakers are using a construction with the same frequency as natives, the distributions on the associated dimension should be the same; conversely, if the distributions differ their prosody may be significantly different. Figure 5 overviews the workflow. The first stages are fully automatic: 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. Thus we applied the loadings discovered by PCA to samples taken every 10 milliseconds throughout the data, simply by taking the dot product.

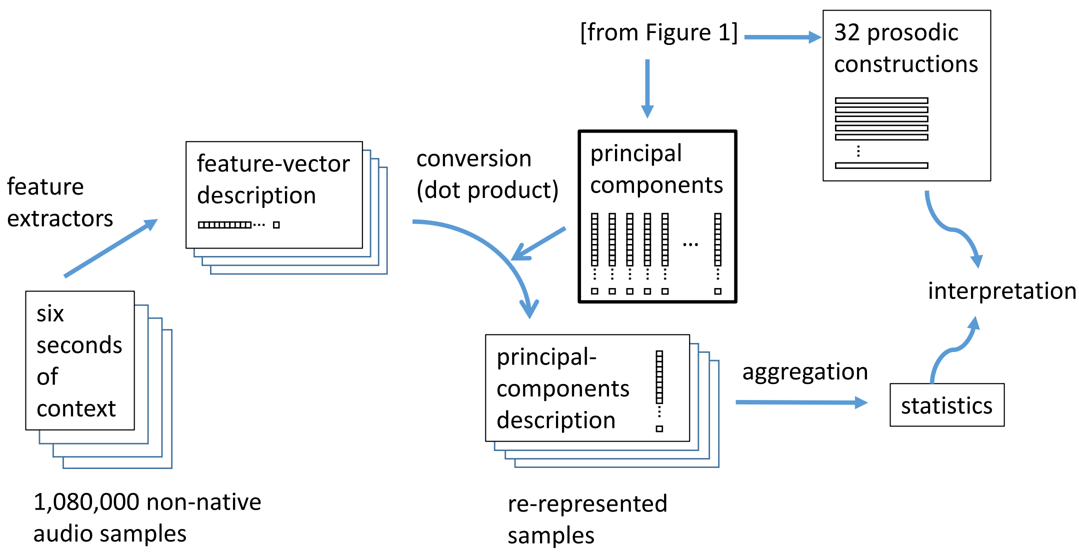

Figure 5: Workflow for comparing distributions.

To test this method, we first applied it to another set of English data to see whether it would find differences that made sense. Specifically we used the Switchboard data.

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. Referring to the interpretation in Table 3, this indicates that in this data less often were both participants simultaneously talking or laughing together. By listening to some conversations we readily confirmed that this was in fact the case. This is unsurprising, given that turn-taking is generally more formal in telephone conversations and in conversations between strangers.

For reasons of space we do not show distributions for the other dimensions. Instead summary statistics are given in Table 5, where columns 2 and 3 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) by the standard deviation of the same dimension in the reference set. Thus in columns 2, 4, and 6 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. In columns 3, 5, and 7, for the standard deviations, values less than 1 mean the Switchboard speakers had narrower distributions than the reference speakers, and greater values wider distributions.

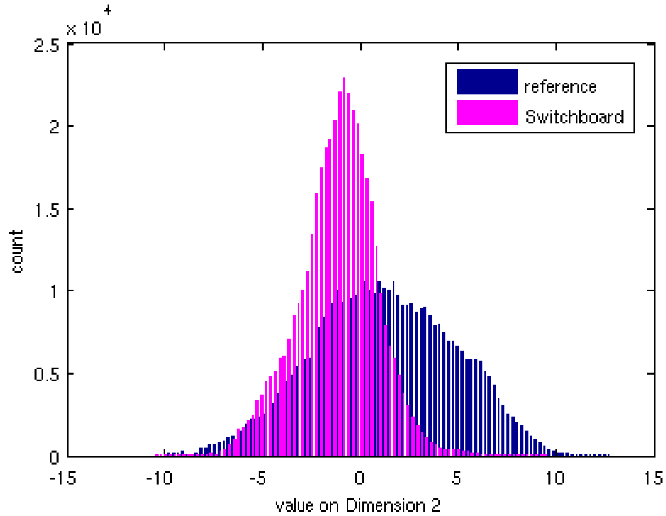

Figure 6: Distribution of values on Dimension 2 for the reference data and the Switchboard data.

Among the differences, the largest was for Dimension 10, for which the Switchboard speakers tended to the negative side. Prosodically, the 10-lo construction involves quieter-than-average utterances, with gradually increasing pitch and a moment of creakier, faster speech with expanded pitch range. Pragmatically this is associated with alignment and agreement, as seen in Example 2. In this fragment G is agreeing that the class R plans to take is interesting, starting quietly and then building, culminating with an enthusiastic *huh*.

Example 2 (eng001@419.2s)

R: but I still really want to do it, because it sounds interesting

G: yeah, that does sound interesting,

R: and (trails off)

G: huh! that'd be interesting ...

Thus the distribution of values on Dimension 10 suggests that Switchboard conversations exhibit more alignment and agreement than the reference dialogs. This was again easily confirmed by some listening. It is also easy to understand: strangers who have no desired outcomes beyond having a pleasant conversation tend to find things they can agree on.

Thus it seems that comparing distributions can reveal differences in prosodic behavior that reflect real differences in dialog activities and interaction styles. Having verified the approach, we next applied the same method to the non-native data, and, for comparison, to the monolingual native data. Table 5 shows the results. Columns 4 through 7 are 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 our native-speaker comparison data. 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 indicates a real difference.

	Switchboard		native		non-native	
dimension	mean	std. dev.	mean	std. dev.	mean	std. dev.
1	-0.00	1.22	-0.00	1.05	**-0.11**	0.97
2	**-0.31**	**0.53**	-0.33	0.76	**-0.45**	**0.75**
3	0.00	0.91	0.00	0.91	0.01	0.86
4	0.00	0.87	-0.00	0.85	**-0.05**	0.83
5	-0.12	0.74	-0.01	0.94	0.01	0.91
6	-0.17	0.65	0.10	0.86	**0.14**	0.84
7	-0.06	0.86	-0.06	0.97	-0.03	0.97
8	-0.00	0.81	0.00	0.86	0.03	0.86
9	0.00	0.85	0.00	0.82	-0.00	0.81
10	**-0.43**	0.97	-0.10	1.01	**-0.21**	0.99
11	-0.00	0.94	-0.00	0.91	-0.01	0.93
12	-0.00	0.92	-0.00	0.89	-0.00	0.89
13	0.11	0.71	0.03	0.91	0.03	0.90
14	**-0.33**	0.79	-0.07	0.94	-0.09	0.95
15	**0.37**	0.87	0.07	0.97	0.07	0.98
16	0.00	0.88	-0.00	0.92	-0.02	0.92
18	-0.00	1.02	-0.00	0.97	**0.11**	0.91
21	0.00	0.95	0.00	0.88	**-0.13**	0.87

Table 5: 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 in columns 2 and 3 marks the largest differences between Switchboard and the reference set, and in columns 6 and 7 the largest differences between the non-native and the native datasets.

However, such differences were slight. Indeed, contrary to expectation, the non-native means were mostly closer to the native means than the Switchboard speakers' means were. We examined the differences statistically, using unmatched, two-tailed heteroskedastic t-tests with Bonferroni corrections, taking as independent samples the means of each speaker's values, and found no statistically significant differences. Nevertheless there appear to be tendencies; the rest of this section discusses the dimensions with larger differences.

For Dimension 1 the averages were noticeably different, indicating that the non-natives speakers talk rather more than the native speakers. This is easy to confirm, by listening, and easy to understand: often language learners take more words and more attempts to convey what they want to say. For Dimension 2 also, the averages were different, indicating that the non-native speakers tended to have less overlapped speech; something that was again easy to confirm by listening.

For Dimension 3, although there was only a tiny difference in means, the non-native data exhibited narrower variation. This dearth of extreme values on this dimension indicates that the non-native speakers had fewer prototypical turn takes and turn yields. From this statistic alone we cannot tell exactly how they differed, but when we listened to the data, we did notice a tendency to have longer gaps between speaker changes.

For Dimension 4 the non-native speakers averaged slightly lower. Dimension 4-hi involved the speaker making an interleaved short contribution in the midst of speech from the interlocutor. These short contributions were usually a backchannel, a short question, laughter, or a suggestion of a word that the other was looking for. (The interlocutor's prosody in the vicinity involved peaks in intensity, creakiness, and enunciation or speaking rate about two seconds apart, often with a gap in between.) This fact that the non-native average was lower indicates that the non-native speakers less commonly produced small utterances precisely interleaved in the other's turn. Again, listening confirmed this tendency.

For Dimension 6 the non-native data averaged slightly higher. The prosody of Dimension 6-hi involves a region of low intensity, generally a pause, surrounded by two regions of high intensity, wide pitch range, and creakiness; thus this indicates a tendency for the non-native speakers to pause more often to think.

For Dimension 10 the non-native speakers averaged lower. As noted above, this indicates a greater tendency to align or agree with the other person. Example 3 is an exemplar of this prosodic pattern: the words *squeeze you* are fast, creaky and in wide pitch range, and they lead into a high-pitched laugh. This example requires some explanation. A and B have been talking about pets, and B has related a time when he tried to use a dog as a pillow. He re-enacts the situation, speaking as he might have addressed the dog in apology. A shows empathy by herself acting out he might have felt, in the last clause herself addressing an imagined dog.

Example 3 (nn007@126.8s)

B: 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 large differences for some others. Space permits discussion of only two.

For Dimension 18 the non-natives speakers had fewer low values, indicating 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. (It is not that the non-native speakers were unfamiliar with wry humor; in fact, when we applied PCA to the Spanish data wry humor showed up quite high, in dimension 10. However the Spanish form is very different, involving a few seconds of creaky voice and a fast speaking rate, including a short period of increased pitch range in the vicinity of a short pause.)

For Dimension 21 the non-native speakers 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.

Thus this method led us to find interesting differences, of various types. Some appear to reflect actual prosodic deficits (Dimension 18 and, as discussed below, Dimension 3). Others appear to reflect processing limitations (1, 3, 4, 6, 21), which is plausible since learners may be slow to comprehend and/or need more time to create fluent utterances (Wiberg, 2003). Yet others seem more to involve cultural factors (2, 10, 21), which are known to often transcend considerations of what language is currently being spoken (Tannen, 1989). For example, the greater use of construction 10-lo (alignment) can be related to well-attested norms of Mexican culture (Condon, 1985). It seems likely that the non-native speakers were behaving as they thought appropriate, rather than trying to behave like native speakers but failing due to a prosodic skill deficit. Similarly, the reduced use of Dimension 21-high (aggressive/rushed turn holding), could be explained as a choice not to use (or not to acquire) a behavior that seems rude in many contexts.

However distributions cannot tell the whole story: there are gross difference in prosodic behavior that they do not show. This point was driven home for us when we looked at the distributions of our *Spanish* data on the same reference dimensions. To our surprise, there were only minor differences in the distributions. We speculate that this is because the space of possible prosodic variation is limited, and so each language tends to use the entire space, although for different purposes. Be that as it may, it was clear that we needed another way to look at the data.

7. Dimension Differences

Mennen's (2015) categories include two other kinds of differences, relating to realization and to semantics. We therefore set out to look for differences in the details of how non-native speakers produce specific prosodic constructions and of how they map pragmatic functions to constructions. To do this we developed a second method, that of comparing the *patterns* of non-native behavior to the native patterns, rather than directly comparing non-native data to the native patterns. Thus, in this approach the first step is to characterize the prosodic behaviors of the non-native speakers in their own terms. This is done, again, using PCA.

Figure 7 shows the concept. We assume that, if the non-native behavior is similar to that of the native behavior in some respect, then the relevant pattern of native behavior will be well matched by some non-native pattern. Conversely, we assume that native patterns that lack a counterpart will be behaviors that the non-native speakers 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 a simple operator for this, the cosine: by this metric, two dimensions are similar if, for every feature, when one dimension's loading on that feature is strongly positive then the other is also, and conversely if negative.

Table 6 shows the result: the cosines between the top five reference dimensions and the top five non-native dimensions. It is clear that the 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 7, 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

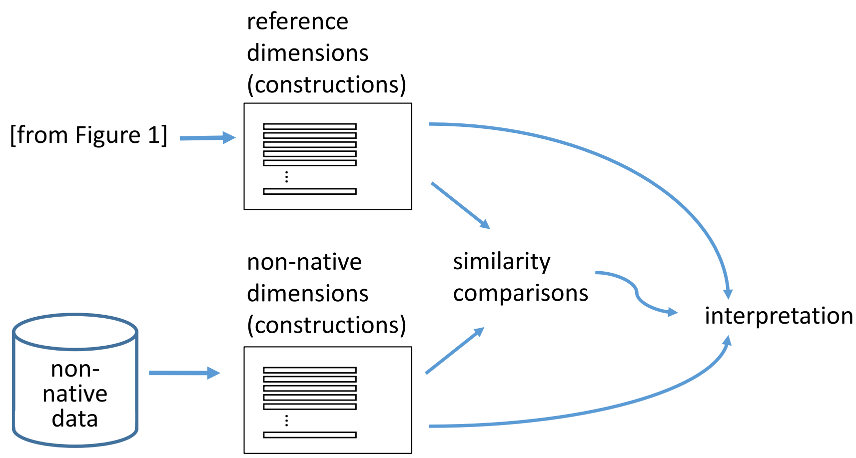

Figure 7: Workflow for comparing construction inventories.

the same time, also unsurprisingly, the natives are almost always closer to the reference than are the non-natives.

We first discuss major differences then minor differences, as they seem to reflect different kinds of issues.

	nn 1	nn 2	nn 3	nn 4	nn 5
ref 1	**.97**	–.07	–.00	–.01	.07
ref 2	.11	**.90**	–.00	.09	–.05
ref 3	–.01	.01	**–.95**	–.10	–.09
ref 4	.04	.03	.08	–.43	**–.80**
ref 5	.03	–.14	.05	**.61**	–.44

Table 6: Similarities between reference dimensions and non-native dimensions. In each row the highest value is bolded.

7.1 Differences in Use

Table 7 indicates that the non-native speakers lack patterns corresponding to the Reference Dimensions 5 and 7. In this subsection we describe the corresponding patterns in native-speaker dialogs, then examine how the non-native speakers differed.

The low-side pattern of Dimension 5 involves about a half-second of increased intensity, starting with high pitch and ending creaky. At times in the native 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. In the non-native data, we found that only some speakers used this prosodic pattern for this function. (As always, we inferred their intended functions from their words and their behavior in the wider context.) One appeared not to use it at all; that is, there were no times

reference dimension	cosine of the most strongly-related dimension native	non-native	ratio
1	.99	.97	.98
2	.90	.90	1.00
3	.92	.95	1.03
4	.92	.80	.87
5	.80	.61	**.76**
6	.64	.63	.98
7	.83	.57	**.69**
8	.73	.63	.86
9	.74	.69	.93
10	.67	.64	.96
11	.87	.83	.95
12	.62	.79	1.27
13	.74	.72	.95
14	.75	.64	.85
15	.74	.68	.92
16	.60	.56	.93

Table 7: For each of the reference dimensions, the cosine of the best-matching dimension found for the other data sets.

where her speech was highly negative on this dimension. Another used this prosody frequently on question-initial *so*, making her questions sound incongruously aggressive.

The high-side pattern of Dimension 5 involves low pitch over several seconds, and within that a short region that is creaky and even more strongly low in pitch. This frequently occurs with words like *and, um, like*, and *you know*, for example when a 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. In the non-native data, while some speakers used this prosodic pattern 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.

For Dimension 7, the low-side pattern involves pitch strongly high and with a region with a fairly slow drop in intensity, rate and creakiness over about 1.5 seconds. Native speakers frequently use this trailing off or "intensity-fade" pattern to solicit empathy, and sometimes also when leaving something unsaid and inviting the listener to infer it. There were many cases where the non-native speakers 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.

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. Listening to places in the non-native data high on Dimension 7, we found no cases where a non-native speaker used a long region 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-native speakers didn't use the long low-pitch region pattern at all; others did, however not for explaining but rather when talking about something personal, such as family background, likes and dislikes, habits, or intentions.

Overall, this suggests that the lack of a non-native dimension corresponding to a reference dimension really does indicate a weakness with their prosodic expression of the corresponding functions

7.2 Differences in Realization

The method is effective not only for identifying gaps in learners' skills, but also for detecting where they are using essentially the same constructions in the same ways, but with small differences. Although small — after all, as Table 7 shows, for many dimensions the non-native differences are minor, reflecting their advanced level — the differences are revealing. In this subsection we consider just the three top dimensions.

Non-native Dimension 1 is very similar in loadings to Reference Dimension 1, as seen by the 0.97 cosine, 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 the loadings are not identical: in particular for some speaking rate features the loadings are higher for the native speakers. This suggests that native speakers talk faster when they have the floor, but the non-native speakers 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 speech by the two speakers, native speakers tend to have a fast speaking rate, whereas non-native speakers tend to have a higher pitch. For Reference Dimension 3 and Non-native Dimension 3, turn hand-offs, the major difference is that the native speakers tend to speak faster at turn starts, but for non-native speakers this tendency is much weaker.

The reality of these differences, suggested by the loadings, was readily confirmed by listening to some of the non-native data. Thus this method also seems valid: the differences that it uncovers are real ones.

However we must note that this method does not appear to be reliable for less-frequent constructions. Looking again at Table 7, 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 speakers appear to be doing this fairly successfully (a cosine of .79), the comparison native speakers appear 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 set of native speakers tended to talk more about technical topics than about personal ones, giving them fewer opportunities to be supportive.

8. Summary and Prospects

We have presented new methods for finding prosodic patterns that are under-used or variant in form in non-native speech. These methods work automatically from data, once an initial analysis of the native-speaker patterns has been done. Software supporting this workflow is available as open-source (Ward, 2015). In these methods we use 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. This improves on previous methods in going beyond mere feature-level differences, while avoiding some of the biases inherent in top-down theory-driven analyses.

We applied these methods to discover some important prosodic constructions of English dialog, and used these to discover ways in which Spanish-native learners differed, including in the prosody of turn-taking, showing involvement, and explaining.

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). We earlier noted that the non-native speakers did not tend to pick up the pace of speaking even when they have the floor (Dimension 1), do not tend to mark involvement (5), tend to agree more (10), and may not mark factual, explanatory information as different from personally-relevant information (7). Thus, while there may be real cultural differences, these cross-cultural perceptions may also reflect mere prosodic-behavior differences.

We note that these results must be interpreted cautiously. Here our aim was to explore new methods and previously unremarked phenomena, not to definitively establish any fact or settle any issue. One limitation of this study was that the speakers were not matched across the conditions, so some of the differences found may reflect different personality types or other uncontrolled differences across the datasets. Another limitation is that the only verifications done involved looking at the data to check the reasonableness of what the automatic methods suggested. A more methodologically-sound procedure would be to first obtain an independently-created and validated listing of non-native dialog-prosody deficits, something that unfortunately does not exist at present. A third limitation is that, in places where the method required subjective judgments, we used only our own. Judgments from more observers would be required to have full confidence in the interpretations given.

There are many open scientific questions relating to our methods and our findings. One is how to find a better feature set, that is, one that leads to results that accurately reflect what is perceptually most salient and communicatively most important. Another set of questions involves the details of each construction. Here we relied heavily on automated methods, seeking a big-picture inventory and a broad-brushstroke understanding. Like many other big-data methods (Swanson and Charniak, 2014) this was efficient, but has its limits. Further examination using more sensitive methods could better tie these construction-based descriptions to those developed within other theoretical frameworks and improve the accuracy of the descriptions.

Another scientific question is which differences in prosody actually 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 hand, such differences may affect perceptions and dialog outcomes (Tannen, 2005; Curhan and Pentland, 2007). Identifying which differences matter will require further study involving broader consideration of interpersonal and social factors.

Despite these limitations and open questions, this work may be useful in various ways.

These findings may be useful for teachers. Second language teaching, when it treats prosody at all, usually focuses on just a few well-understood aspects (Diepenbroek and Derwing, 2013; Busa, 2012). Here we have identified some frequently-used dialog-related aspects of prosody that are almost certainly of value to learners, but are never taught, nor acquired naturally even after years of immersion. Future work should refine our findings into descriptions that will be useful for learners and teachers. Future work should also develop better techniques for teaching the prosody related to interaction patterns in dialog, as this involves special challenges (Ward et al., 2007; Betz and Huth, 2014).

The methods developed here may be useful for rating speakers. Assessment is important for gatekeeping purposes — including placing learners into the appropriate level of instruction – and recently there is growing interest in ways to rate not only language knowledge but also interaction skills and communicative effectiveness. This is true not only for non-native speakers (Mitchell et al., 2014; Litman et al., 2016), but also for native speakers who wish to become more effective in interviews or other unfamiliar situations (Hoque et al., 2013).

Finally, beyond assessment of overall competence, these methods may be useful for pinpointing the prosodic deficits of individual speakers. Among other challenges, this will require ways to obtain reliable results with less data.

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 means of Research Experience for Undergraduates supplements to IIS 191-4868 and IIS 144-9093, and in part by the Fulbright Program.

References

Berit Aronsson and Lars Fant. Boundary tones in non-native speech: The transfer of pragmatics strategies from L1 Swedish into L2 Spanish. *Intercultural Pragmatics*, 11:159–198, 2014.

Amalia Arvaniti. The representation of intonation. In Marc van Oostendorp, Colin J. Ewen, Elizabeth V. Hume, and Keren Rice, editors, *The Blackwell Companion to Phonology*. Wiley, 2011.

Anne-Marie Barraja-Rohan. Using conversation analysis in the second language classroom to teach interactional competence. *Language Teaching Research*, 15:479–507, 2011.

Anne Berry. Spanish and American turn-taking styles: A comparative study. In L. F. Boulton, editor, *Pragmatics and Language Learning Monograph Series, Volume 5*, pages 180–190. University of Illinois, Urbana-Champaign: Division of English as an International Language, 1994.

Emma M. Betz and Thorsten Huth. Beyond grammar: Teaching interaction in the German language classroom. *Die Unterrichtspraxis/Teaching German*, 47(2):140–163, 2014.

Joan Borras-Comes, Constantijn Kaland, Pilar Prieto, and Marc Swerts. Audiovisual correlates of interrogativity: A comparative analysis of Catalan and Dutch. *Journal of Nonverbal Behavior*, 38:53–66, 2014.

J. Donald Bowen. A comparison of the intonation patterns of English and Spanish. *Hispania*, 39:30–35, 1956.

Harry Bunt. Multifunctionality in dialogue. *Computer Speech and Language*, 25:222–245, 2011.

Maria Grazia Busa. The role of prosody in pronunciation teaching: A growing appreciation. In Maria Grazia Busa and Antonio Stella, editors, *Methodological Perspectives on Second Language Prosody*, pages 101–105, 2012.

Gao-Peng Chen, Gérard Bailly, Qing-Feng Liu, and Ren-Hua Wang. A superposed prosodic model for Chinese text-to-speech synthesis. In *International Symposium on Chinese Spoken Language Processing*, pages 177–180. IEEE, 2004.

Zi-He Chen, Yuan-Fu Liao, and Yau-Tarng Juang. Prosody modeling and eigen-prosody analysis for robust speaker recognition. *ICASSP*, 2005.

Herbert H. Clark. *Using Language*. Cambridge University Press, 1996.

John C. Condon. *Good Neighbors: Communicating with the Mexicans*. Intercultural Press, 1985.

Elizabeth Couper-Kuhlen. *An introduction to English prosody*. Edward Arnold, 1986.

Elizabeth Couper-Kuhlen and Margret Selting. *Prosody in Conversation: Interactional Studies*. Cambridge University Press, 1996a.

Elizabeth Couper-Kuhlen and Margret Selting. Towards and interactional perspective on prosody and a prosodic perspective on interaction. In Elizabeth Couper-Kuhlen and Margret Selting, editors, *Prosody in Conversation: Interactional Studies*. Cambridge University Press, 1996b.

Jared R. Curhan and Alex Pentland. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92:802, 2007.

Anne Cutler and D. Robert Ladd. Models and measurements in the study of prosody. In Anne Cutler and D. Robert Ladd, editors, *Prosody: Models and Measurements*. Springer, 1983.

Carme de la Mota, Pedro Martin Butragueno, and Pilar Prieto. Mexican Spanish intonation. In P. Prieto and P. Roseano, editors, *Transcription of Intonation of the Spanish Language*, pages 319–350. Lincom Europa, 2010.

Lori G. Diepenbroek and Tracey M. Derwing. To what extent do popular ESL textbooks incorporate oral fluency and pragmatic development? *TESL Canada Journal*, 30:1–20, 2013.

Maria Estelles-Arguedas. Expressing evidentiality through prosody? prosodic voicing in reported speech in spanish colloquial conversations. *Journal of Pragmatics*, 85:138–154, 2015.

Maria Gabriela Valenzuela Farias. A comparative analysis of intonation between Spanish and English speakers in tag questions, wh-questions, inverted questions, and repetition questions. *Revista Brasileira de Linguística Aplicada*, 13:1061–1083, 2013.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520, 1992.

Adele E. Goldberg. Constructionist approaches. In Thomas Hoffman and Graeme Trousdale, editors, *The Oxford Handbook of Construction Grammar*, pages 15–31. Oxford University Press, 2013.

Agustin Gravano and Julia Hirschberg. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25:601–634, 2011.

Michele Gubian, Lou Boves, and Francesco Cangemi. Joint analysis of F0 and speech rate with functional data analysis. In *ICASSP*, pages 4972–4975, 2011.

John J. Gumperz. *Discourse Strategies*. Cambridge University Press, 1982.

Kieu-Phuong Ha, Samuel Ebner, and Martine Grice. Speech prosody and possible misunderstandings in intercultural talk: A study of listener behaviour in standard Vietnamese and German dialogues. In *Speech Prosody*, pages 801–805, 2016.

Nancy Hedberg, Juan M. Sosa, and Lorna Fadden. The intonation of contradictions in American English. In *Prosody and Pragmatics Conference*, 2003.

Nancy Hedberg, Juan M. Sosa, and Emrah Gorgulu. The meaning of intonation in yes-no questions in American English. *Corpus Linguistics and Linguistic Theory, to appear*, 2014.

John Heritage. Epistemics in action: Action formation and territories of knowledge. *Research on Language & Social Interaction*, 45(1):1–29, 2012.

Julia Hirschberg. Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36:31–43, 2002.

Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing*, pages 697–706, 2013.

Jose Ignacio Hualde. *The Sounds of Spanish, Chapter 14: Intonation*. Cambridge University Press, 2005.

International Standards Organization. Language resource management – semantic annotation framework (SemAF) – part 2: Dialogue acts. ISO 24618-2:2012, 2012.

Shuichi Itahashi and Kimihito Tanaka. A method of classification among Japanese dialects. In *Eurospeech*, 1993.

Oliver Jokisch, Tristan Langenberg, and Gabor Pinter. Intonation-based classification of language proficiency using FDA. In *Speech Prosody*, 2014.

Evia Kainada and Angelos Lengeris. Native language influences on the production of second-language prosody. *Journal of the International Phonetic Association*, 45:269–287, 2015.

Rose Thomas Kalathottukaren, Suzanne C. Purdy, and Elaine Ballard. Behavioral measures to evaluate prosodic skills: A review of assessment tools for children and adults. *Contemporary Issues in Communication Science and Disorders*, 42:138–154, 2015.

Okim Kang. Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, 38:301–315, 2010.

D. Robert Ladd. Stylized intonation. *Language*, pages 517–540, 1978.

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

Catherine Lai. Response types and the prosody of declaratives. In *Speech Prosody*, 2012.

Mark Liberman and Ivan Sag. Prosodic form and discourse function. In *Papers from Tenth Regional Meeting, Chicago Linguistic Society*, pages 402–427, 1974.

Diane Litman, Steve Young, Mark Gales, Kate Knill, Karen Ottewell, Rogier van Dalen, and David Vandyke. Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of English. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–275, 2016.

Ineke Mennen. Beyond segments: Towards a L2 intonation learning theory. In Elisabeth Delais-Roussairie, Mathieu Avanzi, and Sophie Herment, editors, *Prosody and Language in Contact*, pages 171–188. Springer, 2015.

Christopher M. Mitchell, Keelan Evanini, and Klaus Zechner. A trialogue-based spoken dialogue system for assessment of English language learners. In *Proceedings of the International Workshop on Spoken Dialogue Systems (IWSDS)*, 2014.

Oliver Niebuhr. Resistance is futile: The intonation between continuation rise and calling contour in German. In *Interspeech*, pages 132–136, 2014.

Richard Ogden. Prosodies in conversation. In Oliver Niebuhr, editor, *Understanding Prosody: The role of context, function, and communication*, pages 201–217. De Gruyter, 2012.

Richard A. Ogden. Linguistic resources for complaints in conversation. In *International Congress of the Phonetic Sciences*, pages 1321–1324, 2007.

Marta Ortega-Llebaria and Laura Colantoni. L2 English intonation: Relations between form-meaning associations, access to meaning, and L1 transfer. *Studies in Second Language Acquisition*, 36:331–353, 2014.

Benjamin Parrell, Sungbok Lee, and Dani Byrd. Evaluation of prosodic juncture strength using functional data analysis. *Journal of Phonetics*, 41(6):442–452, 2013.

Caterina Petrone and Oliver Niebuhr. On the intonation of German intonation questions: the role of the prenuclear region. *Language and Speech*, 57:108–146, 2013.

Pilar Prieto. Intonational meaning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6: 371–381, 2015.

Dolores Ramirez Verdugo. The nature and patterning of native and non-native intonation in the expression of certainty and uncertainty: Pragmatic effects. *Journal of Pragmatics*, 37: 2086–2115, 2005.

Dolores Ramirez Verdugo. A study of intonation awareness and learning in non-native speakers of English. *Language Awareness*, 15:141–159, 2006.

Ma. Dolores Ramirez Verdugo. Non-native interlanguage intonation systems: A study based on a computerized corpus of Spanish learners of English. *ICAME Journal*, 26:115–132, 2003.

Rajiv Rao. Intonational variation in third party complaints in Spanish. *Journal of Speech Sciences*, 3:141–168, 2013a.

Rajiv Rao. Prosodic consequences of sarcasm versus sincerity in Mexican Spanish. *Concentric: Studies in Linguistics*, 39(2):33–59, 2013b.

Uwe D. Reichel. Linking bottom-up intonation stylization to discourse structure. *Computer Speech and Language*, 28:1340–1365, 2014.

Heidi Riggenbach. Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, 14:423–441, 1991.

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.

Jesus Romero-Trillo. *Pragmatics and prosody in English language teaching*. Springer Science & Business Media, 2012.

Fabian Santiago and Elisabeth Delais-Roussarie. The acquisition of question intonation by Mexican Spanish learners of French. In *Prosody and Language in Contact*, pages 243–270. Springer, 2015.

Bjorn Schuller. Voice and speech analysis in search of states and traits. In Albert Ali Salah and Theo Gevers, editors, *Computer Analysis of Human Behavior*, pages 227–253. Springer, 2011.

Elizabeth E. Shriberg and Andreas Stolcke. Direct modeling of prosody: An overview of applications in automatic speech processing. In *Proceedings of the International Conference on Speech Prosody*, pages 575–582, 2004.

Jack Sidnell. *Conversation analysis: An introduction*. John Wiley & Sons, 2011.

Ben Swanson and Eugene Charniak. Data driven language transfer hypothesis. In *14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 169–173, 2014.

Marc Swerts and S. Zerbian. Intonational differences between L1 and L2 English in South Africa. *Phonetica*, 67:127–146, 2010.

Beatrice Szczepek Reed. *Prosodic orientation in English conversation*. Palgrave, 2006.

Beatrice Szczepek Reed. *Analysing Conversation: An introduction to prosody*. Palgrave Macmillan, 2010.

Beatrice Szczepek Reed. Prosody in conversation: Implications for teaching English pronunciation. In J. Romero-Trillo, editor, *Pragmatics and Prosody in English Language Teaching*, pages 147–168. Springer, 2012.

Deborah Tannen. *That's Not What I Meant! How Conversational Style Makes or Breaks Relationships*. Ballantine, 1989.

Deborah Tannen. Interactional sociolinguistics as a resource for intercultural pragmatics. *Intercultural Pragmatics*, 2:205–208, 2005.

Juhani Toivanen. Tone choice in the English intonation of proficient non-native speakers. In *In Proceedings of Fonetik, Phonum 9*, pages 165–168, 2003.

Jürgen Trouvain and Ulrike Gut. *Non-native prosody: Phonetic description and teaching practice*. Walter de Gruyter, 2007.

Giuseppina Turco, Christine Dimroth, and Bettina Braun. Prosodic and lexical marking of contrast in L2 Italian. *Second Language Research*, 2015.

Kristin J. Van Engen, Melissa Baese-Berk, Rachel E. Baker, Arim Choi, Midam Kim, and Ann R. Bradlow. The Wildcat corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, 53:510–540, 2010.

Jan van Santen, Taniya Mishra, and Esther Klabbers. Prosodic processing. In *Springer Handbook of Speech Processing*, pages 471–488. Springer, 2008.

Jan P.H. van Santen, Taniya Mishra, and Esther Klabbers. Estimating phrase curves in the general superpositional intonation model. In *Fifth ISCA Workshop on Speech Synthesis*, pages 61–66, 2004.

Nigel G. Ward. Automatic discovery of simply-composable prosodic elements. In *Speech Prosody*, pages 915–919, 2014.

Nigel G. Ward. Midlevel prosodic features toolkit. https://github.com/nigelgward/midlevel, 2015.

Nigel G. Ward and Paola Gallardo. A corpus for investigating English-language learners' dialog behaviors. Technical Report UTEP-CS-15-33, University of Texas at El Paso, Department of Computer Science, 2015.

Nigel G. Ward and Alejandro Vega. A bottom-up exploration of the dimensions of dialog state in spoken interaction. In *13th Annual SIGdial Meeting on Discourse and Dialogue*, 2012.

Nigel G. Ward and Steven D. Werner. Data collection for the Similar Segments in Social Speech task. Technical Report UTEP-CS-13-58, University of Texas at El Paso, 2013.

Nigel G. Ward, Rafael Escalante, Yaffa Al Bayyari, and Thamar Solorio. Learning to show you're listening. *Computer Assisted Language Learning*, 20:385–407, 2007.

Nigel G. Ward, Alejandro Vega, and Timo Baumann. Prosodic and temporal features for language modeling for dialog. *Speech Communication*, 54:161–174, 2011.

Nigel G. Ward, Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara. Interactional and pragmatics-related prosodic patterns in Mandarin dialog. In *Speech Prosody*, 2016.

J. C. Wells. *English Intonation: An Introduction*. Cambridge, 2006.

John C. Wells. *Sounds Interesting*. Cambridge, 2014.

Eva Wiberg. Interactional context in L2 dialogues. *Journal of Pragmatics*, 35:389–407, 2003.

Anne Wichmann. Discourse intonation. *Covenant Journal of Language Studies*, 2(1), 2014.

Yi Xu. Speech melody as articulatorily implemented communicative functions. *Speech Communication*, 46:220–251, 2005.

Yi Xu. Speech prosody: A methodological review. *Journal of Speech Sciences*, 1:85–115, 2011.

Frank Zimmerer, Jeanin Jugler, Bistra Andreeva, Bernd Mobius, and Jürgen Trouvain. Too cautious to vary more? A comparison of pitch variation in native and non-native productions of French and German speakers. In *Speech Prosody Conference*, 2014.