Laboratory Phonology
Journal of the Association for Laboratory Phonology

Cole, J and Shattuck-Hufnagel, S 2016 New Methods for Prosodic Transcription: Capturing Variability as a Source of Information. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7(1): 8, pp. 1–29, DOI: http://dx.doi.org/10.5334/labphon.29

JOURNAL ARTICLE

New Methods for Prosodic Transcription: Capturing Variability as a Source of Information

Jennifer Cole[1] and Stefanie Shattuck-Hufnagel[2]

[1] University of Illinois, US

[2] Massachusetts Institute of Technology, US

Corresponding author: Jennifer Cole (jscole@illinois.edu)

Understanding the role of prosody in encoding linguistic meaning and in shaping phonetic form requires the analysis of prosodically annotated speech drawn from a wide variety of speech materials. Yet obtaining accurate and reliable prosodic annotations for even small datasets is challenging due to the time and expertise required. We discuss several factors that make prosodic annotation difficult and impact its reliability, all of which relate to *variability*: in the patterning of prosodic elements (features and structures) as they relate to the linguistic and discourse context, in the acoustic cues for those prosodic elements, and in the parameter values of the cues. We propose two novel methods for prosodic transcription that capture variability as a source of information relevant to the linguistic analysis of prosody. The first is *Rapid Prosody Transcription* (RPT), which can be performed by non-experts using a simple set of unary labels to mark prominence and boundaries based on immediate auditory impression. Inter-transcriber variability is used to calculate continuous-valued prosody 'scores' that are assigned to each word and represent the perceptual salience of its prosodic features or structure. RPT can be used to model the relative influence of top-down factors and acoustic cues in prosody perception, and to model prosodic variation across many dimensions, including language variety, speech style, or speaker's affect. The second proposed method is the identification of individual cues to the contrastive prosodic elements of an utterance. Cue specification provides a link between the contrastive symbolic categories of prosodic structures and the continuous-valued parameters in the acoustic signal, and offers a framework for investigating how factors related to the grammatical and situational context influence the phonetic form of spoken words and phrases. While cue specification as a transcription tool has not yet been explored as RPT has, it has the potential to provide a level of detail that will be useful in modelling systematic context-governed variation in the implementation of prosodic categories, with applications in automatic speech synthesis and recognition, as well as modelling human speech production and perception. We discuss how RPT and cue specification, particularly when combined, can improve the efficiency and reliability of prosodic transcription and how they can be integrated with expert phonological transcription.

1 Introduction

Prosody offers rich data for the investigation of sentence structure, information structure, and pragmatic meaning. Prosody is also an important consideration for models of spoken language processing and discourse interaction. An essential component of any investigation of prosody is a speech sample that is annotated to locate and identify the prosodic *elements* of an utterance—a term we will use throughout this paper to refer to the prosodic structures that locate prominences and boundaries and the tonal or other features associated with those structures. A serious concern for prosody researchers, then, is the difficulty of obtaining prosodic annotation, especially for spontaneous speech that is produced in meaningful communication, where prosodic expression can be rich and varied, but where acoustic variation is generally also the greatest and annotation can be particularly

challenging. This paper begins by considering the reasons why prosodic annotation is difficult despite the expert knowledge and training of the transcriber. The answer appears to be that the prosodic annotation of an utterance is difficult when available cues fail to converge on a unique assignment of prosodic features, taking into account acoustic cues and information in the form of predictions from the syntactic, semantic and discourse context. We propose two innovations that enable prosodic annotation in situations of uncertain or conflicting cues. The proposed methods involve annotation from untrained, non-expert transcribers, and identification of acoustic cues to prosodic features. The proposed methods differ from traditional approaches to prosodic annotation in that they embrace variability in the production and perception of prosody as a source of information. The proposed methods are mutually independent and complement one another, but combined they offer richly detailed information from which we can learn more about the perceptual salience and phonetic expression of contrastive prosodic features, about the role of prosody in conveying meaning, and about the nature of the speech processing systems.

2 Why prosodic annotation is difficult

There are a number of factors that make prosodic annotation challenging. Here we review difficulties related to the acoustic cues to prosodic features, individual speaker differences, and effects of the linguistic context. Some of these same factors are noted by other authors in this collection (Cangemi & Grice, 2016; Frota, 2016; Arvaniti, 2016), to which list Arvaniti also mentions speech rate, style, and task. Cangemi and Grice also point to nuances of pragmatic meaning related to interrogatives and politeness as factors that may condition intonation variation. We also recognize these myriad factors, but do not discuss them further here.

2.1 Reduced and ambiguous acoustic cues

One problem that occurs frequently is that acoustic cues to prosodic features can be reduced or ambiguous. Prosodic cues may be reduced as a consequence of the general phenomenon of phonetic reduction in speaking conditions that favor hypoarticulation, or as the result of overlap or blending of prosodic features in contexts where multiple prosodic features are crowded in a small phonological region (Arvaniti et al., 2006; Grabe et al., 2000; Silverman & Pierrehumbert, 1990). An ambiguous cue is one that appears in a context that is compatible with more than one assignment of prosodic features. For example, lengthened duration of a stressed syllable can occur as a cue to prominence or prosodic phrase boundary; or, an F_0 fall on two consecutive full-vowel syllables can cue either a high tonal prominence on the first syllable (e.g., a H* pitch accent in the ToBI annotation system; see Beckman et al., 2005), or a falling tonal prominence on the second accented syllable (e.g., a H + !H* in the ToBI system).

Faced with reduced or ambiguous cues in the acoustic signal, the transcriber has several options, a point also raised by Cangemi and Grice (2016) in discussing the labeling of variable F_0 contours in Italian. The transcriber may explicitly mark the uncertain presence of a prominence or boundary (e.g., *? to mark an uncertain pitch accent in ToBI annotation); identify alternatives to an uncertain label (e.g., marking "H* *or* L + H*" in a ToBI annotation); or appeal to contextual information to choose the prosodic label that is most likely in the given context (e.g., a prosodic boundary at the location of a syntactic clause boundary). The choice among these options may depend on the nature of the uncertainty, the plausibility of alternative analyses, and the strength of predictions based on contextual information. And though it is useful to allow the transcriber flexibility in choosing the best way to resolve uncertainty in any instance, annotation is slow when transcribers must consider multiple solutions in assigning a prosodic element to given word. The problem is

compounded by the fact that prosodic elements interact, so the prosodic element assigned to a word with ambiguous or uncertain acoustic cues can influence the interpretation of the cues and corresponding prosodic elements of an adjacent word. For example, positing a prosodic phrase boundary at the end of a word has the result that the following word now stands in phrase-initial position, which in turn suggests the presence of phonetic strengthening effects that are often observed for phrase-initial phones. Finally, general criteria for resolving these ambiguities can be difficult to make explicit, thus increasing the likelihood of inter-transcriber disagreement.

2.2 Weighting of acoustic cues

It is widely recognized that prosodic elements are expressed in the acoustic signal through numerous parameters including F_0, overall intensity, spectral tilt, duration, and various measures of voice quality. As with acoustic cues to segmental features, the acoustic cues and cue parameter values associated with an individual prosodic element vary across instances of that element. The patterning of prosodic cue variation has not been widely investigated, so we have much yet to learn about the contribution of individual cues or cue combinations to the perception of prosodic features. Nonetheless, it seems likely that in natural speech settings listeners perceptually integrate cues to identify prosodic elements on the basis of the entire cue package. Ideally, transcribers performing prosodic annotation would also take into consideration the entire cue package when assigning a prosodic feature to a word. Yet the transcriber's task differs from that of ordinary speech perception in that the transcriber customarily receives not only an auditory signal, but also a graphical speech display that visually conveys patterns in F_0, intensity, and formants. Unfortunately, other important cues to prosody, e.g., phone-normalized duration, phonation quality, or spectral tilt, lack a clear visual representation in standard speech displays. This means that information the transcriber receives about which cues are present in the signal is biased, privileging especially F_0 and intensity as cues to prosodic features. There are currently no explicit, objective criteria for weighting acoustic cues to prosodic features, and so we can expect variation among transcribers in cue weighting. Disparities in cue weighting across transcribers can percolate up and result in differences in the assignment of the phonological prosodic elements, compromising comparisons across studies, and thereby diminishing the value of the prosodically annotated speech data.

2.3 Individual speaker differences

Our experience as 'expert' transcribers (i.e., having in-depth familiarity with theoretical claims about the phonological structures and with phonetic correlates of prosody) annotating spontaneous speech materials is that annotation often becomes easier and faster as we gain familiarity with the speech patterns of an individual speaker. This reflects the fact that individual speakers vary in their use of prosody. Production studies of prosody in varieties of British and American English report individual speaker differences in the assignment of prosodic elements marking prominence and boundaries (Grabe, 2004; Yoon, 2010), and in the phonetic encoding of prosodic elements (Cole & Shattuck-Hufnagel, 2011; Peppé et al., 2000). For instance, Peppé et al. (2000) find that British English speakers vary in their use of silent pause and final lengthening as a cue to prosodic phrase boundary in short read-aloud sentences designed to elicit internal juncture ('*cream, buns, and jam*' vs. '*cream buns and jam*'), and Cole and Shattuck-Hufnagel (2011) find similar differences among American English speakers in their use of silent pause to mark a prosodic phrase boundary in utterances imitated from spontaneous speech stimuli. What this means for prosodic annotation is that a transcriber annotating a multi-speaker

database cannot rely on any one acoustic cue as evidence for a particular prosodic feature, though there may be patterns of relative consistency in the prosodic patterns produced by an individual speaker.

2.4 Influence from linguistic context

Prosodic structures and features are assigned to words based in part on their syntactic and semantic context. For example, in English the end of a sentence is often marked with a prosodic phrase boundary, and a pitch accent is very frequently assigned to the rightmost content word (or compound) in eventive sentences, subject to conditions of semantic weight or informativeness. The co-occurrence of prosodic features with specific properties of the syntactic or semantic context enables the listener to predict prosodic elements based on recognized syntactic or semantic properties, prior to (or independent of) consideration of the acoustic cues to prosody. Top-down effects of this sort are demonstrated in a recent study of prosody perception in American English by Cole and her colleagues (Cole, Mo, & Baek, 2010; Cole, Mo, & Hasegawa-Johnson, 2010). In that study, words that are marked as prominent by transcribers (using the RPT method, see Section 4) can be predicted on the basis of acoustic cues from the stressed vowel about as well as they can be predicted from the top-down factors of the word's log-frequency and repeated mention index as measures of informativeness (Cole, Mo, & Hasegawa-Johnson, 2010). These top-down factors make an independent contribution to the prominence prediction model, which suggests that in at least some cases transcribers may mark prominence based on the top-down predictors in the absence of strong cues from the acoustic signal. The same study showed similar effects of partially independent top-down prediction from the syntactic context in transcribers' boundary labeling (Cole, Mo, & Baek, 2010). We reason that these top-down predictions about prosodic elements do not typically trump the acoustic evidence, so that information from the acoustic signal is in fact relevant to the transcriber's choice of labels. In fact, it appears that acoustic cues are primary in prosodic transcription in the default case, as shown by Cole, Mahrt, and Hualde (2014), who find that transcribers assign a higher weight to top-down factors like syntactic context or information status only when explicitly instructed to attend to those factors while performing transcription. In the absence of special instructions, the transcribers' choice of labels is more strongly predicted by acoustic cues; but when acoustic cues are ambiguous, top-down information may play a stronger role.

The influence of top-down processing in predicting prosodic elements poses a problem for prosodic transcription. Even if transcribers are instructed to consider only acoustic cues, and to disregard the syntactic, semantic, and discourse context, as in the ToBI transcription guidelines (Beckman & Ayers Elam, 1997), transcribers can't help but be aware of the linguistic context of a word in a given utterance and they may find it hard to fully suppress predictions stemming from their extensive experience with the prosodic patterns of the target language. Predicted prosodic elements may intrude on the transcriber's evaluation of acoustic cues and may favor annotation of prosodic elements that agree with predictions. Intrusions of this sort may be particularly influential in situations where the acoustic cues are reduced or ambiguous, and may partially explain the challenge of automatic prosodic feature detection based only on acoustic input (see Rosenberg, 2009).

3 Acoustic cues and perceptual criteria in prosodic annotation

There are a number of competing approaches to prosodic annotation. Some approaches annotate prosody using abstract phonological features that characterize distinct levels or types of prominences and boundaries. Two examples of such systems are the Tones and

Break Indices [ToBI] system (Beckman et al., 2005), and the Rhythm and Pitch [RaP] system (Breen et al., 2012). Other approaches use continuous acoustic measures (mainly F_0) that can be automatically calculated from the speech signal. Such systems include the Parametric Intonation Event [PaIntE] model (Möhler & Conkie, 1998), Quantitative Target Approximation [qTA] (Prom-on et al., 2009), and TILT (Taylor, 2000).

Prosodic annotation with abstract phonological elements requires the effort of human transcribers; it is based on the transcriber's auditory impression of prosodic elements related to prominences and boundaries supplemented by visual inspection of the waveform, spectrogram, and F_0 contours of the utterance. The advantage of this method is that the human listener can integrate information from many cues, and can take into account relationships between neighboring prosodic elements, considering, e.g., the degree of juncture between successive boundaries, or the level of prominence across successive pitch accents. These relationships have been shown to be important for the prosodic expression of focus, finality, and certain syntactic dependencies (e.g., Katz & Selkirk, 2011; Ladd, 2008, p. 78; Wagner, 2010). The disadvantage of prosodic annotation using abstract phonological elements is that it is in the end a perceptual annotation, and as such it is at least partly subjective. Labeling criteria must be explicit and very clear in order for annotations to be consistent across transcribers, and must provide a standard measure of reliability, yet currently available labeling guidelines simply do not address the range of annotation challenges that are commonly encountered, and that have been highlighted by several decades of experience with prosodic transcription.

Approaches that use acoustic measures to characterize prosody face different challenges. The acoustic measures themselves can often be extracted without human intervention, though most rely on a prior segmentation of the speech signal to designate the word or syllable from which acoustic measures are taken. If word and syllable segmentation can be done automatically, as in Reichel's (2014) CoPaSaul system, then these methods can be used with little human intervention, avoiding the subjectivity of a perceptual annotation. The shortcoming of these methods is that they do not easily perform the cue integration that is automatic for human listeners, nor do they typically take into account the relationship between the acoustic measures of neighboring (or nearby) words. The point we wish to emphasize here is that the association between a prosodic element and its function in signaling linguistic structure and meaning involves not a single cue, but a constellation of cues, with the cue mixture and cue values subject to variation due to factors such as those described in Section 2.

In a prosodic transcription system using abstract phonological elements (i.e., features and structures) to label prominences and boundaries, the elements serve to bundle the phonetic variables that express the structural and meaning functions of prosody. In other words, abstract prosodic elements play the same mediating role in the mapping between speech and meaning as distinctive features do for the representation of lexical contrast (Jakobson et al., 1952; see Mielke, 2011). Stevens (2002) proposed that, in speech perception, listeners extract individual cues to those features from the signal, and integrate that cue information to determine the feature composition of the speaker's intended words; in more recent work (e.g., Stevens & Keyser, 2010), this proposal was extended to include the addition of individual feature-enhancing cues by speakers. This approach suggests a model for segmental processing based on a distinction between the abstract discrete features that carry meaning functions and the phonetic cues that express those features in the speech signal, with different cue patterns occurring in different contexts. Similarly, prosody can be decomposed into two components, with discrete prosodic elements that encode structural and meaning relations among linguistic units such as words and phrases, phonetic cues that are bundled in systematic and potentially language-specific patterns in

the acoustic speech signal, and different cue patterns in different contexts (such as tonal crowding). We argue here that prosodic transcription must explicitly consider both levels. In distinguishing two levels of prosodic representation—one for abstract elements and one for phonetically detailed cues—our proposal resonates with the proposals by other authors in this collection (Cangemi & Grice, 2016; Frota, 2016; Hualde & Prieto, 2016). Beyond the papers in this collection, our proposal is also somewhat similar to that of Hirst (2005), who posits an abstract feature annotation at a level termed Intonation Function, with a stylized encoding of F_0 target points that define phonetic cues.

Different systems of prosodic transcription propose different sets of labels to mark prosodic elements. These differences may be more than superficial, because the choice of prosodic labels determines how information from the continuous speech signal gets translated into information about linguistic structure (e.g., stress constituents, syntactic or discourse juncture), information structure (e.g, given, new, accessible), or discourse reference (e.g., narrow, contrastive, or corrective focus). The prosodic label set must provide distinct labels for prosodic sound patterns that are associated with different meanings, and for those associated with different structures at the phonological, syntactic, or discourse levels. In other words, a prosodic annotation system (like any sound-level transcription system) embodies a hypothesis about the relationship between the prosodic form of an utterance, in terms of the elements that encode prosody, and the associated function of that prosodic form in conveying linguistic meaning. Frota (2016) makes a similar claim, in stating that "a transcription is an analysis of the intonation system, which ultimately aims to identify the contrastive intonation categories of a given language and establish how they signal meaning." It follows, then, that decisions about the type of labels included in a prosodic annotation system can be made only with reference to the nature of the meaningful linguistic contrasts that are conveyed through prosody.

The necessity of considering meaning distinctions for questions about transcription is not a new claim, and is not specific to prosodic transcription (a point also emphasized by Cangemi & Grice, 2016, and Arvaniti, 2016). Transcription at the level of the phone (a consonant or vowel segment) is always grounded in the criterion of meaningful distinction, including distinctions that carry lexical contrasts, as well as distinctions that characterize systematic positional allophones. Phonetic detail that does not distinguish among meaningful categories may be salient to the listener, it may signal important information (such as speaker identity, her physiological or psychological state, or social affiliations of the speaker), and it may be included in memory representation, but it is not typically represented in speech transcription. Rather, a phone-level transcription is a discretization of the speech signal that labels phones (or other sub-lexical units) with their distinct category labels, e.g., using IPA phone symbols. For purposes of linguistic inquiry, a phonetic transcription that employs discrete, categorical labels is appropriate, since evidence suggests that similarly abstract units are part of the cognitive representation of speech. For instance, recent work on speech perception shows a privileged role for abstract category representations in speech perception (Mitterer & Ernestus, 2008; Ernestus, 2013), and in perceptual learning of the acoustic boundaries between phoneme categories (Cutler, 2008).

Existing approaches to prosodic analysis differ in the degree to which the prosodic labels assigned to words are grounded in a theory of the meaning function of those labels in the language under study. Explicit proposals about the linguistic function of prosodic elements (e.g., pitch accents and boundaries) are found in a number of early works on prosody, including Bolinger (1989), and Gussenhoven (1983), and in some work in the Autosegmental-Metrical framework, including the seminal work on English by Beckman

and Pierrehumbert (1986) and Pierrehumbert and Hirschberg (1990), and later work on German by Grice, Baumann, and Benzmüller (2005) and Baumann and Grice (2006). Studies investigating the meaning function of prosodic elements have been carried out for other languages as well—see papers by numerous authors in Elordieta and Prieto (2012) and in the *Speech Prosody* proceedings volumes beginning in 2002. Yet, while researchers may recognize the importance of meaning criteria in establishing an inventory of prosodic features, prosodic annotation itself is typically carried out without consideration of the meaning function of the prosodic labels assigned to individual words. Indeed, functional considerations may be explicitly set aside so that researchers can use a prosodic annotation to explore the prosodic form-function relationship in a given language. In our view this practice is potentially problematic, a concern also expressed by Arvaniti (2016). In the situation where prosodic annotation is difficult due to ambiguous or unclear cues, as described above, a prosodic label assigned to a word may be of uncertain status at two levels: in its acoustic expression and in its meaning function. Such uncertainty jeopardizes subsequent analysis of the relationship between prosodic elements and their meaning functions, or between prosodic elements and their phonetic expression, and presents a challenge for corpus analyses of spontaneous speech, and more generally for the analysis of any set of speech materials where the prosodic annotation is of uncertain reliability, including analyses of languages where there is little prior work on the inventory of prosodic elements.

4 Prosodic annotation from untrained, non-expert transcribers

We have argued that variability and ambiguity in the acoustic cues to prosody can lead to differences among transcribers in the prosodic labels assigned to a given utterance. In common transcription practice inter-transcriber differences are resolved through consensus, majority vote, or arbitration, to yield a single "true" annotation. In other words inter-transcriber differences are discarded, as noise in the transcription signal; yet those differences can also be viewed as a source of information, e.g., about where ambiguities are likely to arise. This section describes Rapid Prosody Transcription (RPT) as a method that captures inter-transcriber differences in annotation, offering new insight into the interaction between the prosodic form of an utterance and its function in conveying linguistic structural relations and meaning.

RPT is a simple method of transcription in which listeners identify prominences and boundaries, in separate tasks, based on their auditory impression of an utterance. It has been used by Cole and her colleagues in a number of studies investigating prosody in American English (Cole, Mo, & Baek, 2010; Cole, Mo, & Hasegawa-Johnson, 2011; Mo, 2011), Hindi (Jyothi et al., 2014), and Russian (Luchkina & Cole, 2013, 2014). Other work has used RPT for the prosodic analysis of French (Smith, 2011, 2013) and Spanish (Hualde et al., 2016), for the study of L2 English prosody production as judged by L1 listeners (Smith & Edmunds, 2013) and for the study of L2 prosody perception with Japanese learners of English (Pintér et al., 2014).

4.1 RPT method

4.1.1 Transcribers

RPT can be performed by any person with functioning hearing and vision, and though our studies have used transcribers who self-report as having normal hearing and vision, the method could also be used by transcribers with sensory impairment to test the effect of the impairment on prosody perception and comprehension. RPT does not require training in prosodic transcription, or any specific knowledge about prosody, speech, or any aspect

of linguistics. Some knowledge of the target language is deemed necessary in order for the transcriber to be able to pick out words in fluent speech, and transcribers should also have sufficient reading ability to be able to follow a transcript of the speech sample they are listening to. RPT studies have been successfully conducted with native speakers of the target language, with fluent bilinguals, and with language learners at different ability levels. Multiple transcribers from the same speaker/listener population transcribe the same speech materials, and patterns of inter-transcriber agreement are used to calculate the prosody "score" for each word, as described below. We have used 10–22 transcribers with the RPT method in our work, with no noticeable differences in the resulting patterns of agreement related to the number of transcribers.

4.1.2 Speech materials

We have used RPT with speech samples of length varying between 10–60 s in duration, and which typically include one or more syntactic clauses and one or more prosodic phrases. RPT can be used for transcription of speech regardless of genre or style. The emphasis in developing RPT was to have an efficient means of transcribing prosody in spontaneous speech produced in interactive communication contexts, though we have also used it with read-aloud sentences, and for transcription of 'imagined' prosody based only on text materials.

4.1.3 Task

Transcribers listen to recorded speech samples through headphones or speakers, and are asked to mark prominences and boundaries on individual words in a transcript of the speech sample. Transcripts are prepared without punctuation, capitalization, or any other orthographic or font specification that might normally be used to convey prosody in written materials. Transcribers are given minimal instructions, such as (for English): *"mark as prominent words that the speaker has highlighted for the listener, to make them stand out,"* and *"mark boundaries between words that belong to different chunks that serve to group words in a way that helps listeners interpret the utterance."* Depending on the goals of the research, instructions can be varied to draw the transcriber's attention to acoustic cues, to the syntactic or discourse structure properties of utterance, to information structure, discourse context, or any other aspect of the stimulus that can be perceived and judged by the listener (Cole, Hualde, & Mahrt, 2014). Transcribers are given no example transcriptions, and no feedback on their transcription. They are told that listeners may differ in how they perceive the prominences or boundaries for the same utterance, and that such variation is informative for the investigator. The goal is to reduce the transcriber's concern about matching a "correct" transcription. **Figure 1** shows a screen shot of an individual's RPT transcription for an excerpt of spontaneous speech, performed using a customized web interface for digital collection of transcription data (Mahrt, 2015).

A key feature of RPT is that the speech sample is presented to the transcriber in its entirety; the transcriber does not have control over playback and cannot preferentially listen to any portion of the speech sample. RPT can be performed in real-time, as the speech sample plays, in which case the transcriptions reflect the listener's immediate perceptual response as a function of the prior context of the utterance. Practically speaking, real-time transcription is necessary with longer speech samples if listeners are not able to start and stop playback of the recording, since it would otherwise be very difficult to recall prosodic judgments for words that appear early in the sample. We have found it helpful to let listeners hear the speech sample two times for each transcription task, and to allow limited

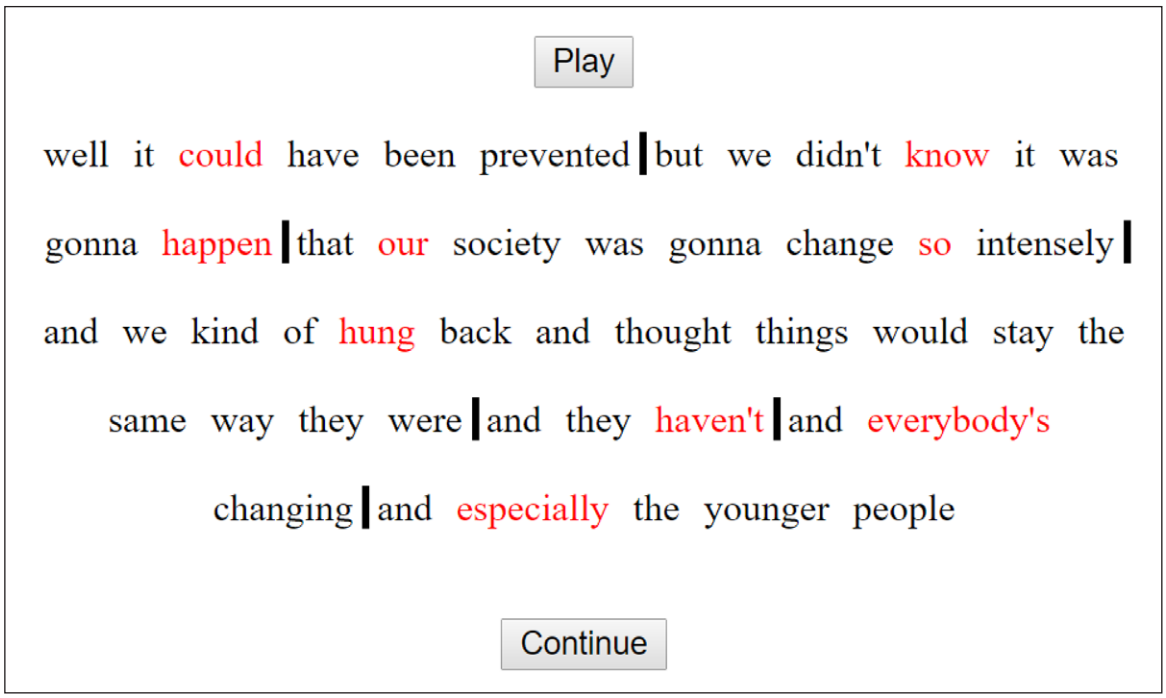

Figure 1: Example RPT output from an individual transcriber showing perceived prominences (red) and boundaries (vertical bars) for a spontaneous speech audio excerpt transcribed in its entirety. Transcript excerpted from the Buckeye Corpus (Pitt et al., 2007), speaker 17, track 1. This audio content is available at: http://dx.doi.org/10.5334/labphon.29.wav1.

modification of labels assigned on the first pass, though within the time constraints of real-time transcription on both listening passes.

Transcription of prominences and boundaries are done as separate tasks when RPT is performed in real-time or rapid mode, because of the difficulty in explicitly attending to these two dimensions of prosody simultaneously, under time pressure. The two tasks can performed by the same transcriber for the same speech materials, either in immediate succession for each speech sample, or sequentially over the entire set of materials transcribed in two blocks of the transcription session, or even in successive sessions.

Transcriptions for each speech sample and all transcribers are aggregated, and each word is assigned a prominence score (p-score) and boundary score (b-score) representing the proportion of transcribers who marked that word as prominent, or as preceding a boundary. **Figure 2** shows the p-scores and b-scores of the words in a sample speech excerpt for American English. Scores near zero indicate that few or no transcribers perceived the word as prominent (or preceding a boundary), while scores near 1 indicate that all or most transcribers perceived the word as prominent (or pre-boundary). Intermediate values represent lesser inter-transcriber agreement.

Figure 2 illustrates patterns we observe frequently in RPT annotations for American English spontaneous speech. First, the most frequent p- and b-score values are zero, indicating that many words in this excerpt are never judged as prominent (8/24) or as preceding a boundary (15/24). Conversely, there are no words that all listeners agree are prominent, and only one word that all heard as preceding a boundary. This finding, which generalizes across our RPT database, indicates that the absence of prosodic marking is more salient than the presence of such marking, suggesting either a clear acoustic "profile" for prosodically unmarked words, and/or a convergence of acoustic cues and top-down factors predicting non-prominence or non-finality. We observe that strong perceptual prosodic breaks (after *though* and *just*) are sparsely distributed, with relatively weaker peaks of

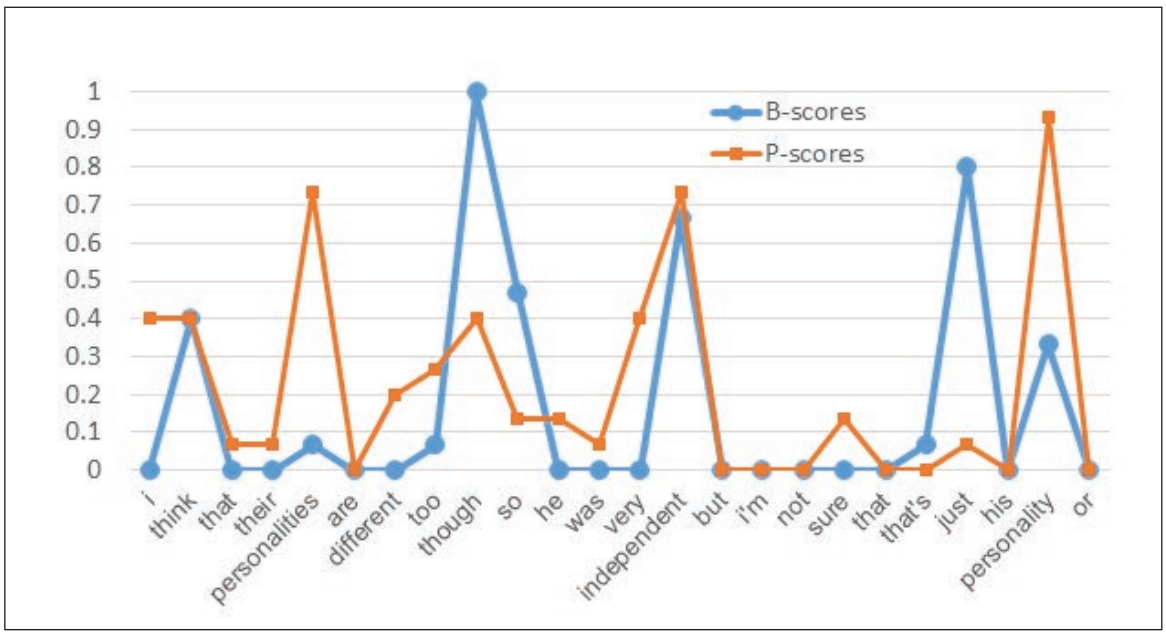

Figure 2: P-scores (prominence) and b-scores (boundaries) for individual words calculated from RPT output from 16 transcribers, for a spontaneous speech sample. Only a portion of the transcribed sample is shown here. Speech sample from the Buckeye Corpus (Pitt et al., 2007), speaker 26, track 1. This audio content is available at: http://dx.doi.org/10.5334/labphon.29.wav2.

perceived prominence on a few words at locations in between. The locations of the strong and weak perceptual breaks suggest a correspondence with the higher and lower prosodic boundaries in hierarchical models of phrase structure (Beckman & Pierrehumbert, 1986; Nespor & Vogel, 1986; Selkirk, 1995). If this finding is confirmed through comparison of RPT with prosodic transcription in one of these frameworks (including the ToBI system), it would indicate a lesser perceptual salience for lower-level prosodic boundaries, which would in turn have implications for our understanding of the role boundaries play in speech processing. Looking at the pattern of perceived prominence across this excerpt, we find less sharply distinguished peaks of prominence, and more numerous smaller prominence peaks, including multi-word spans with diffuse, low-level prominence. These patterns, which again are not unique to this sample, demonstrate that prominence perception is not as consistent across listeners.

4.2 Relating variability in acoustic cues and perception

Although the individual transcriber using RPT makes only binary distinctions for prominence and boundary (marking only those words perceived as prominent or pre-boundary), the p-scores and b-scores derived from RPT are (quasi-)continuous-valued, ranging from 0 to 1, with higher values indicating strong perceptual salience of the prosodic element. This results in graded prosodic labels that can be used to test the contribution of individual acoustic cues or other non-acoustic predictors to the perception of prominence and boundaries. For instance, prior studies show that multiple acoustic measures are correlated with prosodic elements (e.g., Cole et al., 2007; Kochanski et al., 2005), and with the syntactic structures, focus, or information structure that are prosodically encoded (e.g., Breen et al., 2010). The acoustic correlates of prosodic elements are variable across instances of any given prominence or boundary element, a fact well-known to anyone who tries to match training examples to speech samples they are prosodically annotating, which raises the question of whether cue variability contributes to lower inter-transcriber

agreement in prosodic annotation. Is it the case that transcribers are more likely to agree in the assignment of a prosodic label in the presence of strong acoustic cues? The answer appears to be yes, at least for RPT. Several studies that examine acoustic correlates of RPT scores find that higher (i.e., more extreme) cue values predict higher p-scores and b-scores, meaning that transcribers are more likely to mark prominence and boundaries in the presence of strong cues (Cole, Mo, & Baek, 2010; Cole, Mo, & Hasegawa-Johnson, 2010, Mahrt et al., 2012). An interesting question for future research is whether the link between high RPT scores and strong acoustic cues can be used to distinguish among types of prominences or boundaries, or between various structural or meaning functions that those prosodic features encode—e.g., are there stronger acoustic cues signaling a pitch accent marking contrastive focus (e.g., L+H* in the ToBI system) compared to pitch accents marking new-information focus (e.g., ToBI H*), or do stronger cues signal phrases that end at topic boundaries compared to topic-internal phrases? And are RPT p-scores and b-scores also higher for such word tokens with stronger acoustic cues? Preliminary findings for American English are promising (Cole et al., 2014), and we expect that further exploration of these patterns, examining factors that predict the RPT p- and b-scores, may reveal new insights into prosodic elements and their meaning functions.

4.3 RPT and linguistic models of prosody

Like the other papers in this collection, RPT draws on linguistic theories of prosody (or intonation) in recognizing prominence and phrasing as two separate dimensions of prosodic form, and as such RPT can be used within any theoretical framework that recognizes prominence and phrasing, as a means of tapping into ordinary listeners' subjective impression of prominences and boundaries in speech. RPT can even be used to explore prosody, from the perspective of the listener, in languages for which the prosodic phonology has not yet been worked out, and such data may be then used as the basis for developing more articulated grammatical models. For languages where there are existing models of the prosodic phonology, such as English, RPT p-scores and b-scores provide an independent measure of the perceived similarity (or difference) among prosodic elements, such as pitch accents or boundary tones. RPT scores also reveal constraints on the localization and sequencing of prominence and boundaries, in the distribution of scores at different locations in the utterance. For example, a common finding with RPT scores when viewed over individual speech samples, as in **Figure 2**, is that there are diffuse peaks of prominence (or boundaries), where in a sequence of two or more words, each word has an elevated p-score (or b-score), indicating that listeners hear a prosodic event in the region of those words, but that the event is not as clearly localized in perception compared to other locations. In **Figure 2** such an example can be seen in the elevated p-scores over the first two words, …*I think*…, suggesting a prominence on one of those words, though the audio sample for that excerpt does not convey strong emphasis that would be expected if both words were individually and independently prominent.

Another potentially fruitful application of RPT is in research on dialectal variation. Using similar speech materials in related languages, differences in listeners' judgments of prominence and phrasing can be correlated with measurable differences in acoustic cues to gain insight into the parameters of variation. In this collection, the papers by Frota and Prieto and Hualde advocate the adoption of phonetically transparent intonational labels to facilitate comparisons of intonation systems across dialects and languages. RPT can be used independently or as a supplement to this practice, with RPT scores providing an independent measure of the similarity (or differences) in the location of pitch accents or boundary-marking tones across language varieties, as a basis for assigning differences at

the level of the phonological prosodic element, or in the acoustic cues associated with an element.

4.4 Broadening the empirical foundations of prosodic theory

RPT transcription is much faster for the individual transcriber than what is commonly reported for prosodic transcription performed by trained, expert annotators, as in a ToBI transcription. As already mentioned, RPT can be done in real time, and even allowing for two listening passes each for prominence and boundary transcription, RPT can be carried out in something like 4x real time. This is at least one order of magnitude faster than transcription under the ToBI method, as anecdotally reported from many ToBI transcription sites, and consistent with our collective experience. RPT does require many more transcribers than are typically used for an expert transcription, but RPT transcribers can work in parallel. With transcription tasks conducted over the internet, an entire study can be completed in very short time, depending only on the availability of transcribers. The efficiency of RPT at the level of the individual transcriber opens the door for large transcription projects, and for transcription of spontaneous speech or other genres that exhibit high phonetic variability, and including materials with disfluencies or emotional and affective content that can affect the expression of prosodic elements at the level of the acoustic cue.

A further advantage of RPT is that it allows many people to participate as listeners/transcribers in prosody research, including those from populations not easily accessed from the university communities where most prosody researchers reside. This opens the door to obtaining prosody judgments from minority linguistic communities, from elderly people and those in rural communities, and from communities of language learners.

4.5 Prosodic variation across speech genres and styles

An important goal of prosody research is to identify the function of prosody in encoding structural and meaning distinctions. This goal has been pursued in prior work through controlled elicitation of utterances produced in response to prompting questions or statements. For example, Eady et al. (1986) investigate the prosodic encoding of focus in sentences like *Jeff gave the ball to the cat* produced in response to different prompts that elicit broad focus (*What happened?*), focus on the verb phrase (*What did Jeff do?*), or focus on the indirect object (*What did Jeff give the ball to?*). In this case, the information status and focus conditions that are being studied for their relation to prosody are determined in advance by the experimenter and built into the speech materials. The analysis can focus directly on evidence of prosodic elements such as pitch accent, or their absence, on the pre-designated target words. This experimental approach has been successfully used to elicit both read aloud productions (e.g., Eady et al., 1986; Gussenhoven, 1983, and others), and spontaneous productions (e.g., Breen et al., 2010; Speer et al., 2011), and has provided rich insight into the prosodic encoding of information structure, focus, and syntactic structure. But the same experimental methods are not easily extended to investigate prosodic phenomena of a more complex nature, or from a greater variety of discourse contexts.

To the extent that RPT can be used to gauge the perceptual salience of prominences and boundaries in speech representing any genre or style, it may yield insight into the relationship between prosodic elements and their function. For example, radio news announcers tend to have a very lively and engaging speech style that is manifest in the frequent use of pitch accents, relative to conversational speech. This raises the question for English, where pitch accents are used to mark focus, whether listeners associate pitch

accents with information status to the same degree in radio news speech and in conversational speech. Does the frequency of pitch accent in radio news speech lead listeners to discount the value of pitch accents in signaling important information? More generally, do listeners adapt their sensitivity to prosodic cues in the speech signal, or adjust the weighting of prosodically signaled information, as a function of the speaking style, discourse context, or other situational factors? RPT can provide direct answers to such questions, and more generally, offers a window for viewing the prosodic processing in speech.

5 Annotation of acoustic cues

The second approach we propose to advancing prosodic transcription is based on lessons learned from recent work on transcription of segmental variation. These lessons concern the degree and type of surface phonetic variation that spoken utterances exhibit. Over the past few decades, a large body of information has accumulated about the striking degree and systematic nature of context- and situationally-governed variation in the surface phonetic forms of words and their component sounds in spoken utterances (e.g., Bybee, 2001; Hawkins, 2003, 2011; Johnson, 2004; Jurafsky et al., 2001; Kohler, 1998). These observations in the segmental domain have provided a new way to think about transcription. In this section we discuss lessons learned from this work about how the acoustic cues to a given linguistic contrast or structure vary systematically in different contexts, and we examine the potential applicability of these lessons to studies of prosody. Specifically, we suggest that prosodic transcription might benefit from focusing on the question of precisely how the patterns of *individual cues* to the contrastive categories of prominence and boundary types can vary in different contexts. That is, we suggest a focus on the presence vs. absence of individual acoustic cues that are known (or hypothesized) to signal prosodic contrasts, as well as on how both cue choice and cue values vary systematically. We believe that this approach can lead to a better understanding of the contrastive phonemic categories of prosodic structure, as well as of the systematic differences in how these structures are implemented phonetically across contexts, across languages, and across individual speakers.

The identification of individual cues to distinctive features has long been considered in studies of the phonetic implementation of contrastive segmental categories. For example, in a well-known early paper on the cues to the voicing contrast in stop consonants in American English, Lisker (1986) observes that "as many as 16 pattern properties can be counted that may play a role in determining whether a listener reports hearing, for example, *rapid* vs *rabid*" (p. 3). Since then, a large number of authors too numerous to mention here have discussed the active control of phonetic variation in speech processing. For example, Kingston and Diehl (1994) proposed that "the phonetic interpretation of phonological representations may be controlled as well as automatic, because contextual variation in the realization of distinctive feature values is a flexible and adaptive response to variation in the demands on the production or perception of these values between contexts" (p. 419). And more recently, Steriade (1999) and Flemming (2004) have proposed a critical role for the perceptibility of feature cues in the process of phonological and phonetic sound change. Wright (2005) lists a comprehensive set of acoustic cues to segmental feature contrasts, summarizing much of the knowledge in this area that has accumulated over the past 50 or 60 years of speech signal analysis and perceptual experimentation.

While these and many other authors have considered the effects of feature cues on phonological and phonetic processes, the potential implication of separate cognitive representations of cues has not always been thoroughly drawn out. Stevens (2002) took

the further step of proposing distinct roles in perception for different types of feature cues, distinguishing between acoustic **landmarks** (abrupt changes in the acoustic signal, largely associated with manner features), and other acoustic cues (associated with place and voicing features). He proposed that landmark cues are recognized first, providing not only information about manner features, but also the location of regions likely to be rich in other cues (such as formant transitions into and out of consonant closures) and constraints on which kinds of features to search for evidence of, in those landmark-defined regions (e.g., in the region of a vowel landmark, analyses of the signal for cues to the feature [strident] are not required). In work with Keyser (Keyser & Stevens, 2006; Stevens & Keyser, 2010), this proposal for landmark-based initial processing in speech perception was combined with the suggestion that speakers, like listeners, represent individual feature cues, and that they sometimes choose to enhance a feature contrast by adding, strengthening, omitting, or weakening individual feature cues in different contexts. In this approach, a feature cue is not a raw acoustic measure of the speech signal, but rather a derived value, which captures a change in or relationship among measurable values in the speech signal (see below for further discussion).

Traditionally, systematic context-governed phonetic variation has been handled in transcription by the use of allophones, sometimes called positional allophones. However, it is not clear that allophonic transcription is the best way to handle some of the phenomena which have been gradually revealed over the past decades by large-scale analysis of speech corpora made up of speech produced in natural communicative situations. For example, these analyses have revealed many cases of massive reduction (Johnson, 2004), in which, e.g., English words like *totally* and *probably* are produced as something close to (but potentially distinct from) the words *toy* and *pry*. Allophonic transcription, in which each allophonic character represents a symbolic category that summarizes a specific bundle of cues, elegantly captures many aspects of the extensive surface variation in word form that is observed in communicative speech. But it is not well suited to capturing the fact that some of the cues specified by a single allophone relate to one word, phoneme, or feature of the target utterance, while other cues relate to a different word, phoneme, or feature of the utterance—a circumstance that is particularly extensive in cases of massive reduction. In such cases, the transcription might lead one to assume that an entire phoneme had been deleted from the utterance, while in fact one or more cues to the features of an apparently deleted phoneme may well be available in the signal. A processing model in which speakers and listeners represent and manipulate individual cues to feature contrasts as well as their quantitative values provides a natural account of this otherwise puzzling phenomenon. That is, in a cue-based model it is natural that when speakers massively reduce the form of a word, they sometimes leave just a few cues to the target phonemic segments of that word in the signal; this set of remaining cues has been called the 'phonetic residue' (Niebuhr & Kohler, 2011). Such a model is also consistent with a large number of findings illustrating the sensitivity of speakers and listeners to detailed variation in cue values produced by different speakers (such as VOT duration (Neilson, 2011) or the spectral profile of the noise associated with /s/ vs. /f/ (Cutler et al., 2010)).

This proposal for individual-feature-cue-based analysis in the segmental domain has inspired attempts to transcribe individual cues to segmental features in both adult and child speech (Levy et al., 2014; Shattuck-Hufnagel & Veilleux, 2000; Shattuck-Hufnagel et al., 2012; Song et al., 2012). Results have shown that about 80% of the landmark cues predicted by the words of an utterance are implemented as predicted (Shattuck-Hufnagel

& Veilleux, 2002); children sometimes produce non-adult-like cues that enhance feature contrasts (Shattuck-Hufnagel et al., 2015); and careful cue-based analysis of apparently deviant child speech can sometimes help to distinguish a child's typical adult-like contextual variation from a speech disorder (Zhao, 2010).

What are the implications of this individual-feature-cue-based view of segment-level transcription, for the transcription of prosody? If we were to take a parallel feature-cue-based approach to the prosodic domain, then we would take it as the goal of a prosodic transcription system to specify (a) the contrastive phonological elements that define the grammatical prosody of the utterance, i.e., its phrasing and prominence patterns, (b) the acoustic cues to the contrastive features of those prosodic elements, and (c) the values of those cues. This means that we would need to identify (i) the set of abstract symbolic prosodic categories that can distinguish among different structures and meanings, i.e., the phonological contrasts; (ii) the set of acoustic cues that can signal these contrasts in different contexts, as well as the ways speakers choose among these cues; and (iii) the pattern of variation in the signal parameter values associated with those cues. When viewed in this light, the task of resolving the current lack of agreement about the phonological categories of prosodic structure (and their relationship to meaning) takes on new urgency: if there is an analogy between contrastive phonemic categories defined by segmental features and contrastive categories similarly defined by prosodic features, it will be important to determine what these categories and features are (and how they relate to meaning differences), as part of the process of developing a more effective prosodic transcription system.

In the meantime, it is useful to consider what evidence we have that speakers represent and manipulate individual cues to prosodic elements, and that listeners perceive and use these patterns. That is, to what extent might there be a parallel benefit to developing a system for transcribing the individual cues to prosodic elements and their features? And in what ways might such a system be useful in dealing with variation in the surface form of prosodic elements across speakers, contexts, and languages? In the next section we consider briefly some of the individual cues to prosodic boundaries and prominences, knowledge of which has begun to emerge from multiple efforts over the past few decades to specify how prosodic elements and contrasts are signaled, using acoustic measures in corpus-based and experimental studies. We also examine behavioral evidence suggesting that speakers represent individual prosodic cues, and discuss some of the potential benefits of an approach to prosodic transcription that includes cue specification.

5.1 Cues to prosodic elements

For many years, the cues to phrase-level prosody were taken to be F_0 contours, duration adjustments, and amplitude variation. In this traditional view, which pre-dated the proposals (e.g., in Hayes, 1989; Nespor & Vogel, 1986; Selkirk, 1984) for a hierarchy of prosodic constituents and prominences, contrastive levels of stress or prominence were all of the same 'type' and could involve all of these cues, without (for example) a distinction between lexical-level and phrase-level prominence. As more acoustic analyses of prosodic contours have been carried out, it has become clear not only that there are additional cues to prosodic contrasts, but that the choice of cues can vary systematically with context. For example, in the 1990s, Beckman and Edwards (1994) proposed a hierarchy of prominences for English prosody, from reduced syllables to full-vowel syllables to lexically-stressed syllables to phrasally-Pitch-Accented syllables. These authors proposed a different set of cues for each level: e.g., in English, greater duration and amplitude distinguish lexically-stressed syllables from reduced syllables, while F_0 mark-

ers distinguish lexically-stressed syllables from syllables that bear additional phrasal prominence in the form of Pitch Accents. Similarly, it was proposed that speakers also mark the different levels of boundaries in the prosodic constituent hierarchy with different cues and different degrees of, e.g., durational lengthening (Wightman et al., 1992 and Kim et al., 2006 for English; Jun & Fougeron, 2000 for French). Brugos (2015) summarizes a number of cues to prosodic phrase boundaries that have been reported, including "segmental duration, silent intervals between segments, intensity and loudness of the segments, changes to the f0 in which the segments are produced, and voice quality and spectral changes" (p. 21). In particular, she notes two that have not yet been mentioned here: pausing, and pitch reset (sometimes described as the return to a neutral phrase-onset F_0, from a high or low phrase-final F_0 level at the end of the preceding phrase). There is some evidence that boundary cues from pause duration and final lengthening may be interdependent (Fon & Johnson, 2004; Ladd, 1988; Lehiste et al., 1976; Scott, 1982; Prom-On et al., 2009). For example, Fant and Kruckenberg (1989) found evidence in Swedish that if the degree of final lengthening is less, the duration of the following pause is greater.

Other investigators have noted an additional cue to both phrase-level boundaries and prominences: a change in voice-quality. It has long been observed that speakers sometimes produce irregular pitch periods toward the end of an utterance, a phenomenon called final creak. But analyses based on the hierarchy of prosodic constituents revealed that, at least in American English, some speakers also produce irregular pitch periods at the *onsets* of prosodic constituents, and of pitch accented syllables, particularly if those constituents begin with a vowel (or, more rarely, a sonorant consonant). For example, Pierrehumbert and Talkin (1991) showed that speakers often produce an episode of irregular pitch periods at the onset of a phrase-initial vowel, and an analysis of FM radio news speech by Dilley et al. (1996) showed that, for reduced vowels, this is more likely to occur for Full Intonational Phrases than for lower-level Intermediate Intonational Phrases. Dilley et al. also documented the tendency for speakers to produce irregular pitch periods at the onset of vowels that begin a Pitch-Accented syllable, a finding corroborated for additional dialects of English by Garellek (2014). Detailed studies of the articulatory changes that occur at prosodic boundaries have been carried out by Byrd and colleagues (Byrd & Saltzman, 2003; Byrd et al., 2006, Krivokapic & Byrd, 2012), by Krivokapic (2007), and by Katsika et al. (2014), but will not be described in detail here.

The detailed distribution of prosodic prominence- and boundary-related lengthening in the speech signal has been investigated by Turk and White (1999), who studied precisely where accent-related lengthening occurs across the accented syllable and adjacent syllables, and by Turk and Shattuck-Hufnagel (2007), who found that phrase-final lengthening in English was concentrated in the phrase-final rhyme, but also occurred in the rhyme of the main-lexical-stress syllable of the phrase-final word, when the main-stress syllable was not word-final (as in, e.g., *Michigan*).

Over the decades it has been observed that prosodic prominences and boundaries can influence the production of segmental cues as well; for example, Jun (1993) showed that, in Korean, the voice onset time for a stop release increases systematically for preceding boundaries that are higher in the hierarchy of prosodic constituents. Moreover, speakers can adjust the implementation of their prosodic cues to maintain the segmental contrasts of their phonemic system. For example, Nakai and Turk (2011) showed that, in languages with vowel quantity (duration) contrasts, speakers produce less phrase-final lengthening on the phonemically short vowels than on the long vowels, presumably to ensure that the contrast in duration between these two types of vowels is not obscured. Thus it is possible

that systematic variation in cues to segmental features may also provide information to the listener about prosodic structure.

This brief review of some of the types of cues that have been reported for prosodic boundaries and prominences suggests a wide variety of signaling options at the disposal of the speaker (and listener) to signal these prosodic elements. Interestingly, it appears that these cues can sometimes function in a trading relationship, just as has been suggested for cues to segmental features (Repp, 1982). The relation between final duration lengthening and pausing in production, described above, is one example; another is the work of Beach and colleagues (1991; Katz et al., 1996) on cue trading between duration and pitch in the perception of prosodic boundaries. Such functional relationships suggest that speakers can mutually adjust the values of the cues so that they work together to meet the more general goal of signaling a prosodic boundary or prominence. We turn now to some additional evidence that speakers represent and manipulate individual prosodic cues.

5.2 Evidence that speakers represent individual cues

One line of evidence is that speakers can substitute individual cues to prosodic elements, just as they can substitute cues to segmental feature contrasts, e.g., in challenging speaking situations. For example, in whispered Mandarin speech, where the production of canonical pitch cues to lexical tones is not possible, there is evidence that speakers can use amplitude variation to signal the missing pitch contours (Gao, 2002). This might be an example of promoting and exaggerating an existing acoustic correlate of a prosodic feature, since a similar but smaller pattern of amplitude variation appears in typically-phonated versions of the same tone. Another example is found in the behavior of speakers with dysarthria, who have trouble controlling their F_0 contours. Patel (2011) has shown that some dysarthric speakers of English signal the distinction between questions (often produced with rising intonation by typical speakers) and statements (typically produced with falling intonation) by exaggerating the final-syllable duration difference between these two forms. Moreover, interlocutors familiar with the speech of these dysarthric speakers have learned to use these new cues to distinguish the two forms perceptually.

One of the earliest lines of evidence consistent with the hypothesis that speakers and listeners manipulate individual cues and their values came from studies showing that the intonational targets postulated in Autosegmental-Metrical theory (Beckman & Pierrehumbert, 1986; Ladd, 2008; Pierrehumbert, 1980) exhibit systematically different shapes in different contexts (such as in conditions of tonal crowding) or vary across a range of parameter values (Barnes et al., 2012; D'Imperio, 2000; Knight, 2008), and that words and segments take on very different surface phonetic shapes in different prosodic contexts. These differences, sometimes called subphonemic, include such phenomena as hierarchy-related increases in phrase-final lengthening (Wightman et al., 1992). This kind of observation is not explained by the idea that speakers select context-appropriate or positional segmental allophones, because the different degrees of final lengthening are not contrastive in any language, and so don't meet the criterion for allophonic status. While some aspects of contextual variation in surface phonetic form may arise as the more-or-less automatic outcome of conflicting pressures on the production system—e.g., undershoot of an articulatory target due to temporal crowding among competing targets—these observations of systematic subphonemic variation in different prosodic contexts are also consistent with the view that speakers can manipulate cue values, such as the degree of lengthening or the alignment of an F_0 contour with its text, and its scaling, not only to signal contrasting prosodic elements but also in response to

contextual demands. Support for this view is found in the fact that different speakers of the same language may choose different ways of signaling a prosodic element (Peppé et al., 2000), and speakers of different languages may similarly vary in their cue choice and cue settings (Grabe, 2004).

These examples suggest that speakers have the ability to represent and choose among cues to prosodic elements, and to vary their cues to fit circumstances. If this is the case, then the ability to transcribe individual cues might be useful (and perhaps even necessary) in order to study these patterns of cue distribution, and how they vary across different contexts, different speakers, and different languages. If our experience with segmental cue analysis is any guide, there is a considerable amount to be learned from examining these detailed patterns of systematic surface variation. We turn now to an example which illustrates how a prosody transcription approach focused on individual cues might offer some new insights.

5.3 A cue-based system for transcribing prosodic disfluency

An example of the usefulness of cue-based prosodic transcription is found in recent work by Brugos and Shattuck-Hufnagel (2012), who have developed a system for annotating the separate cues to prosodic disfluency. Disfluencies in the form of pauses, lengthenings, and repetitions have long been a challenge to phonologically-based transcription systems like ToBI, which require the listener to determine which well-formed prosodic shape the speaker intended to produce (Beckman & Ayers, 1997). In contrast, the cue-based transcription system proposed by Brugos and Shattuck-Hufnagel draws on work by Arbisi-Kelm (2006), who developed a system to separately label the several phenomena associated with disfluencies produced by speakers who stutter. Rather than labelling a disfluency with a single label, this approach allows the transcriber to specify which set of these individual disfluency-related phenomena have occurred in the disfluency (e.g., pause, lengthening, restart, repetition, editorial remark, etc.) This is particularly useful in light of the fact that the cues can combine in patterns whose distribution is not yet fully understood, e.g., duration lengthening with or without pause and vice versa. The ability to label the individual cues separately opens the possibility of quantifying how often they occur together and in what patterns. Our hypothesis in proposing a cue-based approach to more general aspects of prosodic labelling is that these advantages will also move us toward a better understanding of which cues are used to signal which prosodic elements, and how the cues to these prosodic categories vary in different contexts.

5.4 What would a cue-based transcription system look like?

Given the evidence that speakers and listeners may represent and manipulate individual cues to prosodic structure and prominence, and their quantitative values, how might a cue-based transcription system work? First, there is the problem of defining the cues, and second, the problem of determining how each cue can be annotated. With regard to definitions, as noted earlier, a cue is defined not as a single signal-parameter value, but rather as a relation or pattern among measured values. That is, cues are not simple acoustic measurement points but require some additional processing, to determine, e.g., a change in a measured parameter over time and the direction/size of that change (such as the change in an F_0 contour over time), or a comparison of values in the signal at a given time (such as the energy distribution across the frequency spectrum which specifies spectral peaks). Thus, cues are particularly useful in that they provide a link between measureable acoustic values and contrastive categories. But determining which of the acoustic

correlates of a category actually function as cues for listeners in perception, and which correlates are planned and controlled as cues by the speaker, will require considerable experimentation, just as determining the same kinds of information for segmental feature cues will. As noted above, potential cues to the prosodic elements of boundaries and prominences have been determined to include patterns in the domains of pitch (contours and resets), duration (lengthening and shortening) of both spoken elements and silence, amplitude, voice quality, and changes in cues to segmental categories.

Cues like duration lengthening and F_0 scaling pose an interesting challenge for labelling: what should be the criterion for labelling cues which involve 'more of something than would be expected if there were not a prominence or a boundary here', i.e., which apparently involve comparison with a stored prototypical value. A model based on stored prototypical values (e.g., Beach, 1991) requires a unit for the stored value, and it is not clear what that unit should be. For example, it is unlikely to be the word, since listeners can seemingly recognize prosodic grouping and prominence patterns in utterances consisting of nonwords (e.g., in reiterant speech). One possibility is suggested by Stevens's (2002) proposal that listeners' initial processing of an incoming speech signal involves the detection of acoustic landmarks. As noted earlier, a landmark in Stevens's sense is an abrupt change across a range of frequencies, associated with a change in the manner of articulation, such as a consonant closure or release. The duration of the intervals between such abrupt 'acoustic edges' might provide the initial comparison unit for determining the occurrence of duration lengthening.

Another challenge posed by relational cues such as duration lengthening and shortening is that the same duration value may be interpreted differently depending on the context. For example, a given duration of a rhyme may be perceived as lengthened in an utterance that is spoken rapidly, but as not lengthened in an utterance that is spoken more slowly. While human listeners might be relatively reliable in making this context-based judgment, developing an automatic prosodic cue labeler to take advantage of such information may not be straightforward. One approach would be to update the comparison durations for a particular utterance using the inter-landmark durations mentioned above. Support for this possibility is found in the work of Dilley and Pitt (2010) and Dilley (2015), showing that a given stretch of acoustic material (e.g., the rhyme of *leisure*) can be perceived as part of one word (as in *leisure time*) or of two (as in *leisure or time*), depending on the speaking rate of the preceding material. Additional work by Dilley and McAuley (2008), as well as earlier work by Huss (1978) and others, shows a similarly powerful contextual effect, this time of preceding prominence patterns on the perception of prominence. For example, if the word string *you're all right now* is produced with a H-L-H-L F_0 pattern, and preceded by either a H-L *maybe* or a H-L *perhaps*, after the high-low F_0 pattern on *maybe…* (where the high F_0 signals a High pitch accent on the lexically stressed syllable *may-*), the alternating high-low F_0 pattern on *you're all right now* is more likely to be perceived as signaling a High pitch accent on the high-F_0 words (*you're* and *right*). In contrast, when preceded by a high-low F_0 pattern on *perhaps…* (where the low F_0 signals a Low pitch accent on the lexically stressed syllable *-haps*), the same alternating high-low F_0 pattern on the subsequent words *…you're all right now* is more likely to be perceived as signaling a Low pitch accent on the low F_0 words (*all* and *right*) (Dilley & Shattuck-Hufnagel, 1998, 1999).

Such findings illustrate the fact that labelling of cues to prosodic structure (and the eventual development of tools for the automatic labelling of these cues) is not simply a matter of registering acoustic parameter values, but (like human perception) requires a degree of context- and experience-governed interpretation of evidence for contrastive

phonemic categories. We believe that cue-based labeling, despite its challenges, is a good first step toward the development of a better understanding of the interpretive processes required, both by a model of human prosodic processing, and by an automatic prosody detection algorithm.

5.5 Potential benefits of a cue-based transcription system

Much work will be needed in order to test the hypothesis that these scattered examples represent a more general phenomenon, i.e., that speakers and listeners represent and manipulate individual cues to prosodic elements, as well as the parameter values of those cues. However, to the extent that this hypothesis is correct, a cue-based approach to transcription will offer substantial advantages, by capturing the aspects of utterances that speakers and listeners are attending to and representing in speech processing. Transcriptions in terms of individual cues also offers the possibility of enabling the development of an automatic prosody transcription algorithm, since cues can be specified in quantitative signal-processing terms. The trade-off, however, is that this approach requires more complex integration and interpretation of the cues as evidence for the categories. It must be noted here that prosody is particularly complex to transcribe, in any approach, in part because its acoustic correlates (including traditionally-defined patterns of duration, F_0, intensity and, as increasingly recognized, voice quality) are influenced not only by the grammatical categories of phrasing and prominence, but also by speaker- and situation-specific factors, such as the speaker's emotional state and the social relationship between interlocutors. (This comes as no surprise, since the past few decades of research on phonetic behavior has revealed that these same complex factors influence the production of cues to segmental features as well—but the degree to which structure and social-emotional factors are signaled by overlapping cues appears to be greater in prosody, and thus more challenging to unravel.) Untangling the influences of these factors in a way that facilitates more effective use of prosody in automatic speech recognition and synthesis, as well as more complete models of human speech perception and production, may well be encouraged by the development of a prosodic transcription system focused on contrastive categories, their individual cues, and the variation of these cues (and the range of acoustic values they can take on) across contexts, speakers, and languages, with the flexibility to tailor the transcription level to the research task at hand.

6 Conclusion

Given the significance of prosody in conveying grammatical and extra-grammatical meaning in spoken language, establishing a system for identifying and characterizing the prosodic elements in speech is a research priority. We have described challenges that arise in the use of existing systems for prosodic transcription due to variability in the prosodic encoding of grammatical information, in the individual acoustic cues that express prosodic elements, and in the acoustic cue values. We have also argued, along with Arvaniti (2016), that advancing prosodic transcription toward the goal of greater accuracy and reliability ultimately depends on having a better understanding of how prosody functions to mark grammatical contrasts. A further requirement is to understand how prosodic elements and their acoustic cue implementation vary according to context, such as speech style or other situational factors. Along with Cangemi and Grice (2016), we see a critical need for research on variability in phonetic prosodic cues in relation to their meaning functions, and in relation to their phonological categories.

We propose two novel methods in prosodic transcription, which differ from current approaches in their handling of variability. The first method, Rapid Prosody Transcription

(RPT), records listeners' immediate impressions of prosodic elements and explicitly captures information about the perceptual salience of those elements, by calculating prominence and boundary 'scores' for each word as real-number values based on inter-transcriber agreement. The prosody scores simultaneously reflect all factors that influence prosody perception, including acoustic cues and top-down factors, and reveal the conditions in which ambiguity arises in the phonetic expression of prosodic elements. The patterning of these prosody scores may shed light on prosodic encoding of grammatical information at different levels of linguistic organization (e.g., word, phrase, utterance, discourse segment), and on the mechanisms for prosody processing in speech production and perception.

The second method proposed here is the identification of individual cues to the contrastive prosodic elements of an utterance. This focus on individual cues has the potential to provide a link between the contrastive symbolic categories of prosodic structures and their signal parameters, as well as a framework for investigating the systematic effects of a wide range of contextual factors, both grammatical and situational, on the surface phonetic forms of spoken words.

RPT and cue specification are complementary methods that we believe are most informative when used together, or in conjunction with expert transcription that is grounded in a linguistic model of the contrastive prosodic elements or their function in conveying structural or meaning distinctions. We believe that, combined, these methods can yield transcriptions that inform us about the patterning of contrastive prosodic elements as an aspect of phonological form, the perception of those prosodic elements, and the variable expression of those elements in individual acoustic cues and cue parameter values.

At first glance, it may seem that performing RPT with cue specification alongside expert phonological transcription will greatly increase the effort required for prosodic transcription, but in our experience this is not necessarily the case. As for RPT, it is an efficient transcription method because it does not require trained experts and can be performed in real time (or double) for each of prominence and boundary annotation. RPT scores are useful input for a detailed expert transcription because they provide a kind of perceptual weighting of prosodic elements, pointing out areas of perceptual ambiguity as well as areas where prosodic elements are clearer. This information allows transcribers to choose where to allocate their attention, for example, by prioritizing labeling of words with prosody scores that indicate high inter-transcriber agreement as prominent or adjacent to a boundary. Cue specification, on the other hand, can be done through a combination of quantitative and qualitative methods. Quantitative measures are most appropriate for cues that are not easily rendered in visual format on the graphical speech display, such as durational lengthening or shortening. Other cues, such as the occurrence of irregular pitch periods (IPP), can be identified qualitatively based on visual evidence from the waveform and/or spectrogram, and trained transcribers can therefore record the presence of such cues in a manual cue-level annotation. Our experience suggests that manual transcription of cues like IPP may be more straightforward than phonological transcription because cue annotation hews more closely to the acoustic signal, thereby minimizing the difficult perceptual judgments that often arise in transcribing higher-level phonological constructs such as pitch accent. Indeed, we anticipate advances in automatic methods for detection of all cues, minimizing or eliminating the involvement of an expert human transcriber at this level. Cue specification could be used in conjunction with other information from the syntactic or discourse context and the perceived meaning to induce phonological categories in an empirically grounded and rigorous fashion. Furthermore, allowing that in some research scenarios human transcribers remain necessary, as for example with a prosodic

transcription using ToBI, information about individual cues and their acoustic parameter values can aid the expert human transcriber by providing explicit criteria for assignment of a prosodic label. To summarize, the resources required for RPT and cue specification are likely to lead to a savings in the effort required for expert transcription, while also opening new avenues for automated transcription.

Depending on the goals of the individual researcher, RPT and cue specification may be used to augment or facilitate prosodic annotation performed by trained experts, or as noted above, to induce the phonological prosodic categories for a given language, offering a new avenue for prosodic analysis of previously undescribed languages. We believe that RPT and cue specification are especially important for research that considers the perceptual processing of prosody, and by extension, the role of prosody in speech-mediated social interactions. Looking forward, we think the use of RPT and the identification of individual prosodic cues will enrich prosodic transcription by capturing information about variability, while also bringing gains in efficiency. Adopting these practices into the transcription toolkit will provide richer data from a wider range of speech materials that will help in constructing models of prosodic elements and their function in speech communication, across languages.

Supplementary Files

For accompanying TextGrid, Pitch, and wav files, go to http://dx.doi.org/10.5334/labphon.29.smo

Acknowledgements

We have benefited from hearing many perspectives on prosodic transcription from participants at the first workshop on Advancing Prosodic Transcription in Stuttgart, 2012 and from the authors of papers in this special collection. SSH gratefully acknowledges the intellectual basis of the cue-based approach to transcription drawn from the work of Kenneth Stevens on cues to segmental features, and the support provided by MIT's Undergraduate Research Opportunities Program for the development of training materials for LM labelling. We thank two anonymous reviewers and the editors for helpful comments on an earlier version of our paper, and assign them no responsibility for remaining shortcomings. The research reported here was made possible in part with NSF grants BCS 12-51343 and BCS 12-51134 to the first author.

Competing Interests

The authors declare that they have no competing interests.

References

Arbisi-Kelm, T. 2006. *An intonational analysis of disfluency patterns in stuttering* (Unpublished Ph.D. dissertation). Department of Linguistics, University of California at Los Angeles.

Arvaniti, A. 2016. Analytical Decisions in Intonation Research and the Role of Representations: Lessons from Romani. *Laboratory Phonology: Journal of the Association for Laboratory Phonology,* 7(1): 6, pp. 1–43, DOI: http://dx.doi.org/10.5334/labphon.14

Arvaniti, A., Ladd, D. R., & Mennen, I. 2006. Phonetic effects of focus and 'tonal crowding' in intonation: evidence from Greek polar questions. *Speech Communication, 48*, 667–696. DOI: http://dx.doi.org/10.1016/j.specom.2005.09.012

Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. 2012. Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Journal of Laboratory Phonology, 3*(2), 337–383. DOI: http://dx.doi.org/10.1515/lp-2012-0017

Baumann, S., & Grice, M. 2006.The Intonation of accessibility. *Journal of Pragmatics, 38*(10), 1636–1657. DOI: http://dx.doi.org/10.1016/j.pragma.2005.03.017

Beach, C. M. 1991. The interpretation of prosodic patterns at points of syntactic structure ambiguity: evidence for cue trading relationships. *Journal of Memory and Language, 30*, 644–663. DOI: http://dx.doi.org/10.1016/0749-596X(91)90030-N

Beckman, M., & Edwards, J. 1994. Articulatory evidence for differentiating stress categories. In Keating, P. A. (Ed.), *Papers in Laboratory Phonology III: Phonological structure and phonetic form*. Cambridge, UK: Cambridge University Press, pp. 7–33.

Beckman, M. E., & Elam, G. A. 1997. *Guidelines for ToBI labelling, v3*. The Ohio State University Research Foundation. Retrieved from http://www.ling.ohio-state.edu/~tobi/ame_tobi/.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. 2005. The original ToBI system and the evolution of the ToBI framework. In Jun, S.-A. (Ed.), *Prosodic typology: The phonology of intonation and phrasing*. Oxford, UK: Oxford University Press, pp. 9–54. DOI: http://dx.doi.org/10.1093/acprof:oso/9780199249633.003.0002

Beckman, M., & Pierrehumbert, J. 1986. Intonational structure in Japanese and English. *Phonology Yearbook*, *3*, 255–309. DOI: http://dx.doi.org/10.1017/S095267570000066X

Bolinger, D. 1989. *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press.

Breen, M., Dilley, L. C., Kraemer, J., & Gibson, E. 2012. Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory, 8*(2), 277–312. DOI: http://dx.doi.org/10.1515/cllt-2012-0011

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. 2010. Acoustic correlates of information structure. *Language and Cognitive Processes, 25*(7), 1044–1098. DOI: http://dx.doi.org/10.1080/01690965.2010.504378

Brugos, A. 2015. *The interaction of pitch and timing in the perception of prosodic grouping* (Unpublished Ph.D. dissertation). Department of Applied Linguistics, Boston University

Brugos, A., & Shattuck-Hufnagel, S. 2012. A proposal for labelling prosodic disfluencies in ToBI. Poster presented at Advancing Prosodic Transcription for Spoken Language Science and Technology, July 31, 2012, Stuttgart, Germany. Retrieved from http://blogs.bu.edu/prosodylab/publications/.

Bybee, J. 2001. *Phonology and language use*. Cambridge: Cambridge University Press. DOI: http://dx.doi.org/10.1017/CBO9780511612886

Byrd, D., Krivokapic, J., & Lee, S. 2006. How far, how long: On the temporal scope of phrase boundary effects. *Journal of the Acoustical Society of America, 120*, 1589–1599. DOI: http://dx.doi.org/10.1121/1.2217135

Byrd, D., & Saltzman, E. 2003. The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics, 31*, 149–180. DOI: http://dx.doi.org/10.1016/S0095-4470(02)00085-2

Cangemi, F., & Grice, M. 2016. The Importance of a Distributional Approach to Categoriality in Autosegmental-Metrical Accounts of Intonation. *Laboratory Phonology: Journal of the Association for Laboratory Phonology,* 7(1): 9, pp. 1–20, DOI: http://dx.doi.org/10.5334/labphon.28

Cole, J., Hualde, J. I., Mahrt, T., Eager, C., & Im, S. 2014. *The perception of phrasal prominence in conversational speech.* Poster presented at Laboratory Phonology 14, Tokyo.

Cole, J., Kim, H., Choi, H., & Hasegawa-Johnson, M. 2007. Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics, 35*, 180–209. DOI: http://dx.doi.org/10.1016/j.wocn.2006.03.004

Cole, J., Mahrt, T., & Hualde, J. I. 2014. Listening for sound, listening for meaning: Task effects on prosodic transcription. *Proceedings of Speech Prosody 7*, Dublin.

Cole, J., Mo, Y., & Baek, S. 2010. The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes, 25*(7), 1141–1177. DOI: http://dx.doi.org/10.1080/01690960903525507

Cole, J., Mo, Y., & Hasegawa-Johnson, M. 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology, 1*, 425–452. DOI: http://dx.doi.org/10.1515/labphon.2010.022

Cole, J., & Shattuck-Hufnagel, S. (2011). The phonology and phonetics of perceived prosody: What do listeners imitate? *Proceedings of Interspeech* 2011, Florence, Italy.

Cutler, A. 2008. The abstract representations in speech processing. *The Quarterly Journal of Experimental Psychology, 61*(11), 1601–1619. DOI: http://dx.doi.org/10.1080/13803390802218542

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. 2010. How abstract phonemic categories are necessary for coping with speaker-related variation. In Fougeron, C., Kühnert, B., D'Imperio, M., & Vallée, N. (Eds.), *Laboratory Phonology 10*. Berlin: de Gruyter, pp. 91–111.

Dilley, L., & McAuley, D. 2008. Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language, 59*, 294–311. DOI: http://dx.doi.org/10.1016/j.jml.2008.06.006

Dilley, L., & Pitt, M. 2010. Altering context speech rate can cause words to appear or disappear. *Psychological Science, 21*(11), 1664–1670. DOI: http://dx.doi.org/10.1177/0956797610384743

Dilley, L., Pitt, M., Szostak, C., & Baese-Berk, M. 2015. Rate-dependent processing can be speech-specific: Evidence from the disappearance of words under changes in context speech rate. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: University of Glasgow. ISBN 978-0-85261-941-4. Paper number 0915.1-5. Retrieved from http://www.icphs2015.info/pdfs/Papers/ICPHS0915.pdf.

Dilley, L., & Shattuck-Hufnagel, S. 1998. Ambiguity in prominence perception in spoken utterances of American English. In *Proceedings of the 16th International Congress on Acoustics and 135th Meeting of the Acoustical Society of America.* DOI: http://dx.doi.org/10.1121/1.421799

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics, 24*, 423–444. DOI: http://dx.doi.org/10.1006/jpho.1996.0023

D'Imperio, M. 2000. The role of perception in defining tonal targets and their alignment (Unpublished Ph.D. dissertation). Department of Linguistics, The Ohio State University.

Eady, S. J., Cooper, W. E., Klouda, G. V., Müller, P. R., & Lotts, D. W. 1986. Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments. *Language and Speech, 29*(3), 233–251.

Elordieta, G., & Prieto, P. (Eds.). 2012. *Prosody and meaning*. Vol. 25. Berlin: Walter de Gruyter. DOI: http://dx.doi.org/10.1515/9783110261790

Ernestus, M. 2013. Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua, 142*, 27–41.

Fant, G., & Kruckenberg, A. 1989. Preliminaries to the study of Swedish prose reading and reading style. *Speech Transmission Laboratory-QPSR, 2*, 1–83.

Flemming, E. 2004. Contrast and perceptual distinctiveness. In Hayes, B., Kirchner, R., & Steriade, D., (Eds.), *Phonetically-based phonology*. Cambridge, England: Cambridge University Press, pp. 232–276. DOI: http://dx.doi.org/10.1017/CBO9780511486401.008

Fon, J., & Johnson, K. 2004. Syllable onset intervals as an indicator of discourse and syntactic boundaries in Tawan Mandarin. *Language and Speech, 47*(1), 57–82. DOI: http://dx.doi.org/10.1177/00238309040470010301

Frota, S. 2016. Surface and Structure: Transcribing Intonation within and across Languages. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 7(1): 7, pp. 1–19. DOI: http://dx.doi.org/10.5334/labphon.10

Gao, M. 2002. *Tones in whispered Chinese: articulatory features and perceptual cues* (Unpublished Ph.D. dissertation). University of Victoria.

Garellek, M. 2014. Voice quality strengthening and glottalization. *Journal of Phonetics, 45*, 106–113. DOI: http://dx.doi.org/10.1016/j.wocn.2014.04.001

Grabe, E. 2004. Intonational variation in urban dialects of English spoken in the British Isles. In Gilles, p., & Peters, J. (Eds.), *Regional variation in intonation*. Linguistische Arbeiten. Tuebingen: Niemeyer, pp. 9–31.

Grabe, E., Post, B., Nolan, F., & Farrar, K. 2000. Pitch accent realization in four varieties of British English. *Journal of Phonetics, 28*(2), 161–185. DOI: http://dx.doi.org/10.006/jpho.2000.0111

Grice, M., Baumann, S., & Benzmuller, R. 2005. German intonation in autosegmental-metrical phonology. In Jun, S.-A. (Ed.), *Prosodic typology: The phonology of intonation and phrasing*. Oxford: Oxford University Press, pp. 55–83. DOI: http://dx.doi.org/10.1093/acprof:oso/9780199249633.003.0003

Gussenhoven, C. 1983. Testing the reality of focus domains. *Language and Speech, 26*, 61–80.

Hawkins, S. 2003. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics, 31*, 373–405. DOI: http://dx.doi.org/10.1016/j.wocn.2003.09.006

Hawkins, S. 2011. Does phonetic detail guide situation-specific speech recognition? In *Proceedings of the 17th International Congress of Phonetic Sciences*, pp. 9–18.

Hayes, B. 1989. The prosodic hierarchy in meter. In Kiparsky, P., & Youmans, G. (Eds.), *Phonetics and phonology 1: Rhythm and meter,* pp. 201–260. DOI: http://dx.doi.org/10.1016/b978-0-12-409340-9.50013-9

Hirst, D. J. 2005. Form and function in the representation of speech prosody. *Speech Communication, 46*(3), 334–347. DOI: http://dx.doi.org/10.1016/j.specom.2005.02.020

Hualde, J. I., Cole, J., Smith, C., Eager, C., Mahrt, T., & Napoleão de Souza, R. 2016. The perception of phrasal prominence in English, Spanish and French conversational speech. *Proceedings of Speech Prosody 8,* Boston.

Hualde, J. I. and Prieto, P. 2016. Towards an International Prosodic Alphabet (IPrA). *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 7(1): 5, pp. 1–25. DOI: http://dx.doi.org/10.5334/labphon.11

Jakobson, R., Fant, G. M., & Halle, M. 1952. *Preliminaries to speech analysis: The distinctive features and their correlates.* Cambridge, MA: MIT Press.

Johnson, K. 2004. Massive reduction in conversational American English. In Yoneyama, K., & Maekawa, K. (Eds.) *Spontaneous speech: Data and analysis. Proceedings of the 1st Session of the 10th International Symposium.* Tokyo, Japan: The National International Institute for Japanese Language, pp. 29–54.

Jun, S.-A. 1993. *The phonetics and phonology of Korean prosody* (Ph.D. dissertation). Ohio State University (published 1996 by Garland Publishing Inc., New York).

Jun, S.-A., & Fougeron, C. 2000. A phonological model of French intonation. In Botinis, A. (Ed.), *Intonation: Analysis, modeling and technology*. Dordrecht: Kluwer Academic Publishers, pp. 209–242. DOI: http://dx.doi.org/10.1007/978-94-011-4317-2_10

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. 2001. The effect of language model probability on pronunciation reduction. In *Proceedings of ICASSP-01* II.801-804, Salt Lake City, Utah. DOI: http://dx.doi.org/10.1109/icassp.2001.941036

Jyothi, P., Cole, J., Hasegawa-Johnson, M., & Puri, V. 2014. An investigation of prosody in Hindi narrative speech. *Proceedings of Speech Prosody 7*, Dublin.

Katsika, A., Shattuck-Hufnagel, S., Mooshammer, C., Tiede, M., & Goldstein, L. 2014. Effects of compatible vs. competing rhythic grouping on errors and timing variability in speech. *Language and Speech, 57*(4), 544–562. DOI: http://dx.doi.org/10.1177/0023830913512776

Katz, W., Beach, C., Jenouri, K., & Verma, S. 1996. Duration and F0 correlates of phrase boundaries in productions by children and adults. *Journal of the Acoustical Society of America*, *99*, 3179–3191. DOI: http://dx.doi.org/10.1121/1.414802

Katz, J., & Selkirk, E. 2011. Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. *Language, 87*, 771–816. DOI: http://dx.doi.org/10.1353/lan.2011.0076

Keyser, S. J., & Stevens, K. N. 2006. Enhancement and overlap in the speech chain. *Language, 82*(1), 33–63. DOI: http://dx.doi.org/10.1353/lan.2006.0051

Kim, H., Yoon, T., Cole, J., & Hasegawa-Johnson, M. 2006. Acoustic differentiation of L- and L-L% in Switchboard and Radio News speech. *Proceedings of Speech Prosody* 2006, Dresden.

Knight, R.-A. 2008. The shape of nuclear falls and their effect on the perception of pitch and prominence: Peaks vs. plateaux. *Language and Speech, 51*(3), 223–244. DOI: http://dx.doi.org/10.1177/0023830908098541

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. 2005. Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America, 118*(2), 1038–1054. DOI: http://dx.doi.org/10.1121/1.1923349

Kohler, K. 1998. The disappearance of words in connected speech. *ZAS Working Papers in Linguistics*, *11*, 21–34.

Krivokapic, J. 2007. Prosodic planning: effects of phrasal length and complexity on pause duration. *Journal of Phonetics, 35*, 162–179. DOI: http://dx.doi.org/10.1016/j.wocn.2006.04.001

Krivokapic, J., & Byrd, D. 2012. Prosodic boundary strength: An articulatory and perceptual study. *Journal of Phonetics, 40*, 430–442. DOI: http://dx.doi.org/10.1016/j.wocn.2012.02.011

Ladd, D. R. 1988. Declination and 'reset' and the hierarchical organization of utterances. *Journal of the Acoustical Society of America, 84*(2), 530–544. DOI: http://dx.doi.org/10.1121/1.396830

Ladd, D. R. 2008. *Intonational phonology* (2nd ed.). Cambridge, UK and New York, NY: Cambridge University Press. DOI: http://dx.doi.org/10.1017/CBO9780511808814

Lehiste, I., Olive, J. P., & Streeter, L. A. 1976. The role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America, 60*(5), 1–4. DOI: http://dx.doi.org/10.1121/1.381180

Levy, C., Mann, A., Kenney, J., Choi, J.-Y., & Shattuck-Hufnagel, S. 2014. Contextual landmark analysis of speech from typically and atypically developing children. *Journal*

of the Acoustical Society of America, 135, 2293. DOI: http://dx.doi.org/10.1121/1.4877531

Lisker, L. "Voicing" in English: a catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech, 29*(1), 3–11.

Luchkina, T., & Cole, J. 2013. Routes to prominence in free word order language discourse. In *Proceedings of the Workshop on Prosody-Discourse Interface,* Leuven, Belgium.

Luchkina, T., & Cole, J. 2014. Structural and prosodic correlates of prominence in free word order language discourse. *Proceedings of Speech Prosody 7,* Dublin.

Mahrt, T. 2015. Language markup and experimental design software (LMEDS). Retrieved from http://prosody.beckman.illinois.edu/lmeds.html.

Mahrt, T., Cole, J., Fleck, M., & Hasegawa-Johnson, M. 2012. Modeling speaker variation in cues to prominence using the Bayesian information criterion. *Proceedings of Speech Prosody 6*, Shanghai.

Mielke, J. 2011. Distinctive features. In van Oostendorp, M., Ewen, C. J., Hume, E., & Rice, K. (Eds),*The Blackwell companion to phonology*. Blackwell Publishing, Blackwell Reference Online. Retrieved from http://www.companiontophonology.com/subscriber/tocnode?id = g9781405184236_chunk_g978140518423619.

Mitterer, H., & Ernestus, M. 2008. The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109*(1), 168–173. DOI: http://dx.doi.org/10.1016/j.cognition.2008.08.002

Mo, Y. 2011. *Prosody production and perception with conversational speech* (Unpublished Ph.D. dissertation). Department of Linguistics, University of Illinois.

Möhler, G., & Conkie, A. 1998. Parametric modeling of intonation using vector quantization. *Proceedings of the Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.

Nakai, S., & Turk, A. 2011. Separability of prosodic phrase boundary and phonemic information. *Journal of the Acoustical Society of America, 129*(2), 966–976. DOI: http://dx.doi.org/10.1121/1.3514419

Neilson, K. 2011. Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39*, 132–142. DOI: http://dx.doi.org/10.1016/j.wocn.2010.12.007

Nespor, M., & Vogel, I. 1986. *Prosodic phonology*. Dordrecht: Foris. Republished 2007, Berlin: Mouton de Gruyter.

Niebuhr, O., & Kohler, K. J. 2011. Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics, 39*(3), 319–329. DOI: http://dx.doi.org/10.1016/j.wocn.2010.12.003

Patel, R. 2011. Acoustic characteristics of the question-statement contrast in severe dysarthria due to cerebral palsy. *Journal of Speech, Language and Hearing Research, 46*, 1401–1415. DOI: http://dx.doi.org/10.1044/1092-4388(2003/109)

Peppé, S., Maxim, J., & Wells, B. 2000. Prosodic variation in southern British English. *Language and Speech,* 43(3), 309–334. DOI: http://dx.doi.org/10.1177/00238309000430030501

Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation* (Unpublished Ph.D. thesis). Department of Linguistics and Philosophy, Massachusetts Institute of Technology.

Pierrehumbert, J., & Hirschberg, J. 1990. The meaning of intonational contours in the interpretation of discourse. In Cohen, P., Morgan, J., & Pollack, M. (Eds.), *Intentions in communication*. Cambridge, MA: MIT Press, pp. 271–311.

Pierrehumbert, J., & Talkin, D. 1991. Lenition of /h/ and glottal stop. *Papers in Laboratory Phonology II*. Cambridge, UK: Cambridge University Press, pp. 90–117.

Pintér, G., Mizuguchi, S., & Yamato, K. 2014. Boundary and prominence perception by Japanese learners of English: A preliminary study. *Phonological Studies, 17*, 59–66.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. 2007. Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Prom-On, S., Xu, Y., & Thipakorn, B. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of the Acoustical Society of America, 125*(1), 405–424. DOI: http://dx.doi.org/10.1121/1.3037222

Reichel, U. D. 2014. Linking bottom-up intonation stylization to discourse structure. *Computer Speech and Language*, 28, 1340–1365. DOI: http://dx.doi.org/10.1016/j.csl.2014.03.005

Repp, B. H. 1982. Phonetic and auditory trading relations between acoustic cues in speech perception: preliminary results. *Haskins Laboratories Status Report on Speech Research* SR-67/68.

Rosenberg, A. 2009. *Automatic detection and classification of prosodic events* (Unpublished Ph.D. dissertation). Columbia University.

Scott, D. 1982. Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America, 71*(4), 996–1007. DOI: http://dx.doi.org/10.1121/1.387581

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

Selkirk, E. O. 1995. Sentence prosody: Intonation, stress, and phrasing. In Goldsmith, J. (Ed.), *The handbook of phonological theory*. London: Blackwell, pp. 550–569.

Shattuck-Hufnagel, S., Hanson, H., & Zhao, S. 2015. Feature-cue-based processing of speech: A developmental perspective. *Proceedings of the International Congress of Phonetic Science*, Glasgow, Great Britain.

Shattuck-Hufnagel, S., & Veilleux, N. 2000. The special phonological characteristics of monosyllabic function words in American English. *Proceedings of the International Conference on Spoken Language Processing*, Beijing, People's Republic of China.

Silverman, K., & Pierrehumbert, J. 1990. The timing of prenuclear high accents in English. In Kingston, J., & Beckman, M. E. (Eds.), *Papers in Laboratory Phonology 1: Between the grammar and the physics of speech*. Cambridge, UK: Cambridge University Press, pp. 72–106. DOI: http://dx.doi.org/10.1017/CBO9780511627736.005

Smith, C. 2011. Perception of prominence and boundaries by naïve French listeners. *Proceedings of the XVIIth International Congress of Phonetic Sciences*, Hong Kong, pp. 1874–1877.

Smith, C. 2013. French listeners' perceptions of prominence and phrasing are differentially affected by instruction set. *Proceedings of Meetings on Acoustics*, *19*(1), 60191. DOI: http://dx.doi.org/10.1121/1.4799041

Smith, C., & Edmunds, P. 2013. Native listeners' perception of prosody in L1 and L2 reading. *Proceedings of Interspeech*, Lyon, France, 235–238.

Song, J.-Y., Demuth, K., & Shattuck-Hufnagel, S. 2012. The development of acoustic cues to coda contrasts in young children learning American English. *Journal of the Acoustical Society of America, 131*(4), 3036–3050. DOI: http://dx.doi.org/10.1121/1.3687467

Speer, S. R., Warren, P., & Schafer, A. J. 2011. Situationally independent prosodic phrasing. *Laboratory Phonology, 2*(1), 35–98. DOI: http://dx.doi.org/10.1515/labphon.2011.002

Steriade, D. 1999. Phonetics in phonology: The case of laryngeal neutralization. In Gordon, M. (Ed.), *Papers in phonology*. *UCLA Working Papers in Linguistics, 2*, 25–146. Los Angeles: Department of Linguistics, UCLA

Stevens, K. N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America, 111*(4), 1872–1891. DOI: http://dx.doi.org/10.1121/1.1458026

Stevens, K. N., & Keyser, S. J. 2010. Quantal theory, enhancement and overlap. *Journal of Phonetics, 38*, 10–19. DOI: http://dx.doi.org/10.1016/j.wocn.2008.10.004

Taylor, P. 2000. Analysis and synthesis of intonation using the tilt model. *The Journal of the Acoustical Society of America, 107*(3), 1697–1714. DOI: http://dx.doi.org/10.1121/1.428453

Turk, A. E., & Shattuck-Hufnagel, S. 2007. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics, 35*(4), 445–472. DOI: http://dx.doi.org/10.1016/j.wocn.2006.12.001

Turk, A., & White, L. 1999. Structural influences on accentual lengthening. *Journal of Phonetics, 27*, 171–206. DOI: http://dx.doi.org/10.1006/jpho.1999.0093

Wagner, M. 2010. Prosody and recursion in coordinate structures and beyond. *Natural Language and Linguistic Theory, 28*, 183–237. DOI: http://dx.doi.org/10.1007/s11049-009-9086-0

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91*(3), 1707–1717. DOI: http://dx.doi.org/10.1121/1.402450

Wright, R. 2005. A review of perceptual cues and cue robustness. In Hayes, B., Kirchner, R., & Steriade, D. (Eds.), *Phonetically-based phonology*. Cambridge: Cambridge University Press, pp. 34–57.

Yoon, T.-J. 2010. Speaker consistency in the realization of prosodic prominence in the Boston University Radio Speech Corpus. *Proceedings of Speech Prosody,* Chicago, Illinois.

Zhao, S. 2010. Stop-like modification of the dental fricative /dh/: An acoustic analysis. *Journal of the Acoustical Society of America, 128*(4), 2009–2020. DOI: http://dx.doi.org/10.1121/1.3478856

How to cite this article: Cole, J and Shattuck-Hufnagel, S 2016 New Methods for Prosodic Transcription: Capturing Variability as a Source of Information. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7(1):8, pp.1–29, DOI: http://dx.doi.org/10.5334/labphon.29

Published: 30 June 2016