An Investigation of Prosody in Hindi Narrative Speech

Preethi Jyothi[1], *Jennifer Cole*[1,2], *Mark Hasegawa-Johnson*[1,3], *Vandana Puri*

[1]Beckman Institute, University of Illinois at Urbana-Champaign, USA
[2]Department of Linguistics, University of Illinois at Urbana-Champaign, USA
[3]Department of ECE, University of Illinois at Urbana-Champaign, USA
{pjyothi,jscole,jhasegaw}@illinois.edu,vanu.p.sharma@gmail.com

Abstract

This paper investigates how prosodic elements such as prominences and prosodic boundaries in Hindi are perceived. We approach this using data from three sources: (i) native speakers of Hindi without any linguistic expertise (ii) a linguistically trained expert in Hindi prosody and finally, (iii) classifiers trained on English for automatic prominence and boundary detection. We use speech from a corpus of Hindi narrative speech for our experiments. Our results indicate that non-expert transcribers do not have a consistent notion of prosodic prominences. However, they show considerable agreement regarding the placement of prosodic boundaries. Also, relative to the non-expert transcribers, there is higher agreement between the expert transcriber and the automatically derived labels for prominence (and prosodic boundaries); this suggests the possibility of using classifiers for the automatic prediction of these prosodic events in Hindi.

Index Terms: Hindi prosody, perception study, automatic labeling of prosodic events in Hindi.

1. Introduction

Hindi is one of the most widely spoken Indo-European languages in the world with over 200 million native speakers in northern parts of India. There have been a sizeable number of studies on intonation in Hindi. Many early works studied the phenomenon of lexical stress in Hindi words which manifests itself as prominence via a designated syllable [1, 2, 3, 4] and acoustic evidence for lexical stress in Hindi [1, 2, 4, 5]. There are two consistent observations regarding prosody in Hindi across previous work (refer to [1, 3, 6, 7] among others): 1) every content word (i.e. a prosodic word), except for the phrase-final one, is associated with a rising pitch contour and 2) focus induces post-focal pitch range compression.

Prior work confirms the presence of pitch accents on content words with a rising pitch contour. However, there are varying opinions regarding the reason these pitch contours are triggered [8, 9, 10]. Most recently, Féry and colleagues [6, 9, 11] claim that Hindi does not have prominence-lending pitch accents and uses only prosodic phrasing to structure an utterance, with edge-marking phrase tones. As evidence for this claim, they note that Hindi speakers do not produce a consistent pattern of pitch movement on a stressed syllable, and in general have very weak intuitions about the location of stress prominence at the word level.

A question of particular interest to us, and one of the objectives of this paper, is to investigate whether ordinary untrained native listeners of Hindi consistently perceive prosodic elements in Hindi speech (specifically, prosodic prominence and prosodic phrase boundaries). This technique of involving ordinary listeners to derive prosodic transcriptions (the latter will henceforth be referred to as *non-expert transcriptions*) was successfully implemented by Cole et al. [12] for English. This is categorically different from most of the previous work on Hindi prosody; the latter is predominantly based on production studies where trained experts analyzed sentences spoken by native Hindi speakers for evidence of various prosodic elements. To our knowledge, there has not been a systematic enquiry into ordinary listeners' perception of prosody in Hindi speech. We use a corpus of Hindi narrative speech to conduct our perception study, described in more detail in Section 2.

This paper also attempts to initiate the discussion about whether we can automatically detect prosodic elements such as pitch accents and prosodic boundaries in Hindi speech. There is a large body of research that studies the identification and classification of prosodic events in English ([13, 14, 15, 16, 17] are a sampling of some of the important works in automatic labeling of English prosody). Automatic prosody labeling is a relatively unexplored area for Hindi. Many of these studies make heavy use of the ToBI Standard [18] – a formalized notation developed to describe the intonation of Standard American English. Recently, a publicly available toolkit called AuToBI [19] has been developed to automatically detect and classify prosodic events (using ToBI labels) in English. As a first step, we use models of prosody trained on English obtained via AuToBI to automatically label prosodic pitch accents and phrase boundaries in Hindi speech. We hope this investigation informs us of what would be needed to build improved models of Hindi prosody. This could also prove to be useful for the design of automatic speech recognition systems in Hindi.

To summarize, the objectives of this paper are two-fold:

1. *Perception of prosody in Hindi by ordinary listeners*: What is the untrained, ordinary listener's perception of prosodic prominence and phrase boundaries in Hindi? How does this compare to the prosodic transcription by a linguistically trained expert Hindi listener? Do native listeners consistently identify pitch accents in Hindi speech? What about phrase boundaries? These are some of the questions we try to address; the experiments are detailed in Section 3.

2. *Automatic labeling of prosody in Hindi*: How do trained models of prosody in English perform when evaluated on Hindi data? Is the automatic labeling of prosodic events more consistent with the non-expert transcriptions or the expert transcription? What can be deduced from the Hindi evaluation task to build better prosody models for Hindi? These questions are discussed further in Section 4.

We conclude this paper with a closing discussion along with scope for future work in Section 5.

Figure 1: A screenshot of the user interface for the study. Prosodically prominent words turn red on being selected.

2. Speech materials

The speech material used in our experiments was drawn from the "OGI Multi-language Telephone Speech Corpus" [20]. This corpus consists of telephone speech in eleven languages, including Hindi; the corpus has recorded speech from 198 Hindi speakers. The Hindi speech was collected in narrative form by asking volunteers to talk about any topic for up to a minute. Sixty-eight of these one-minute audio clips have corresponding hand-labeled phonetic transcriptions.

Out of the sixty-eight audio clips with phonetic transcripts, we selected ten and extracted excerpts, one from each clip, averaging 24.10 secs in length and averaging 59.2 in the number of words per excerpt; there are a total of 592 words over all excerpts. Since the speech in the OGI corpus was collected from volunteers speaking impromptu, it contains many occurrences of conversational elements such as disfluencies, hesitations and repetitions. Our ten excerpts were chosen such that the utterances in each excerpt were relatively free of disfluencies and the usage of English words were kept to a minimum.[1]

3. Perception of prosody in Hindi by non-expert transcribers

3.1. Method and Experimental Setup

Ten adult native speakers of Hindi participated in this study.[2] All the participants were English-speaking students living in the United States. Most of them could speak and write only Hindi and English (and not other languages); three of them could additionally understand other Indian languages such as Kannada, Oriya and Bhojpuri. Information about their language background was retrieved via a questionnaire administered along with the main study.

The entire study was conducted with the help of a web-based software [21]. The interface of the experiment and the instructions for the prosody transcription tasks were worded in Hindi. This was done in order to, hopefully, make the participants more receptive to prosodic elements in the Hindi excerpts.

The non-expert transcribers were shown the ten excerpts described in Section 2 in a randomized order. For each audio file, a participant was asked to complete two tasks – firstly, listen to how the speaker breaks up the text into chunks and mark the location of chunk boundaries (*prosodic phrase boundaries*) and secondly, mark the words that are emphasized or stand out relative to the other words in the utterance (*prosodic prominence*). The participants were explicitly informed that the phrase chunks do not necessarily have to coincide with any punctuation. In order to get acquainted with the two tasks, the experiment was preceded by a training session using one speech excerpt.

On identifying a chunk, the participant was asked to select the final word in the chunk and a "/" delimiter was inserted after the word to mark the phrase boundary. For the second task of identifying emphasized words, the participants' boundary markers from the previous task were kept visible for them to refer to. Figure 1 shows a snapshot of the interface for an excerpt during the prominence marking task. There was no limitation on the response times. The entire experiment was set up such that it would not exceed an hour. However, the participants could listen to each excerpt any number of times and they could choose to devote any amount of time to each excerpt.

As a second set of experiments, this entire run of two tasks per excerpt for all ten excerpts was repeated – only this time the transcripts were displayed without any accompanying audio clip. Each participant was asked to 'listen' to their own inner speech while reading the excerpt text and mark the word chunks and emphasized words. By removing the associated speech clips, the participants would have to rely entirely on lexico-syntactic cues to guide their annotation.

Finally, we also asked an expert in Hindi prosody to transcribe the same data using ToBI.[3] The expert transcriber was provided the audio signals, along with pitch and intensity tracks and phonetic and word alignments (the annotations were done using the Praat [22] toolkit). We note here that the conditions under which the expert interprets prosody is positively different from the conditions for ordinary listeners. Our experimental results also point to this difference, as detailed in Section 3.2.

3.2. Experimental results and discussion

We first compute Cohen's kappa agreement coefficients [23] between the ordinary listeners' and the expert's transcriptions of prominences and boundaries. This is a fairly standard measure of agreement that takes into account the chance probability of agreement. Values of 0.01 to 0.2 indicate slight agreement, 0.21 to 0.4 is fair agreement and 0.41 to 0.6 indicates moderate agreement. Fig. 2 shows the distribution of agreement coefficients across all the participants for both prominence and boundary labels. This shows that ordinary listeners perceive boundaries (with a moderate $\kappa = 0.41$) much more similarly to the expert than prominences (with a slight $\kappa = 0.15$). The slight agreement for prominence marking in Hindi has a parallel in English. In English, prominences that are more closely tied to meaning (e.g., the nuclear prominence that marks focus) are more reliably marked by transcribers than pre-nuclear prominences, which may serve a rhythmic function ([24, 25]).

We also compute Fleiss' kappa statistics (typically used to compute agreement across multiple transcribers) across all listeners (ignoring the expert transcriptions). Fig. 3 shows Fleiss' coefficients for both prominences and boundaries; *With audio* and *Without audio* specify non-expert transcriptions obtained

[1] Most of the volunteers who helped collect Hindi data for the OGI multi-language corpus were graduate students in the United States and often made use of English words while speaking impromptu in Hindi.

[2] One participant listed "Marwari" as their native tongue but identifies themselves as a native Hindi speaker.

[3] The last author served as our expert. She is a native speaker of Hindi and a simultaneous English-Hindi bilingual (much like the non-expert transcribers).

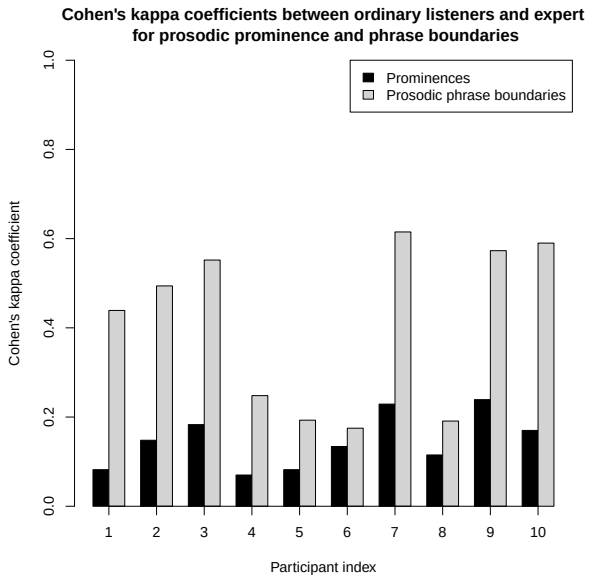

Figure 2: Cohen's kappa agreement coefficients for each participant against the expert for prominences and boundaries.

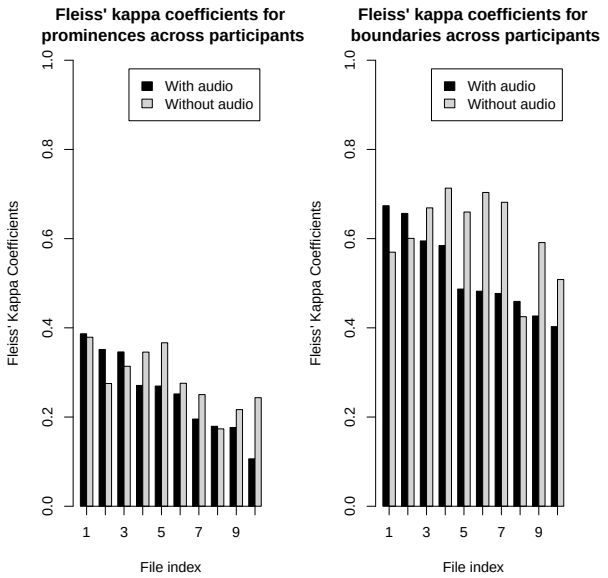

Figure 3: Fleiss' agreement statistics for prosodic prominences and boundaries across all the participants.

with audio and without audio, respectively.[4] We focus first on the non-expert transcriptions with audio. We observe that the non-expert transcribers agree on the location of boundaries well above chance (mean $\kappa = 0.524$) and agree with one another more than they agree with the expert ($\kappa = 0.407$). There is only a fair amount of agreement on prominences ($\kappa = 0.253$)). This partially supports Féry's [9] prediction that Hindi speakers will reliably and consistently perceive prosodic phrases in Hindi utterances and will not reliably perceive prosodic prominence as distinct from phrasing in Hindi utterances.

Comparing the non-expert transcriptions with and without audio, the mean κ values for both prominences and boundaries are comparable (0.25 vs. 0.28 and 0.52 vs. 0.61, respectively). The distributions of non-expert transcriptions with and without audio in Fig. 3, however, suggest that the transcribers may not be getting cues from the same information structure.

Finally, we compute the rate of occurrence of prominences and boundaries. The mean length of intervals (in words) between prominences range from 5.4 – 11.4 for each audio clip, across all transcribers. This indicates the speaker dependent variation of the rate of prominences. Similarly, for boundaries, this range is 5.3 – 8.4. We also compute the mean prominence and boundary intervals for each listener averaged over data across all the clips: 5.0 – 14.0 and 4.6 – 18.5, respectively. This corresponds to listener dependent variation. We note that the listener dependent variation is larger than the speaker dependent variation as previously observed for American English [26].

4. Automatic prosody detection in Hindi

4.1. Method and Experimental Setup

AuToBI [19] is a publicly available toolkit to automatically detect the presence and type of prosodic events, from the ToBI standard, present in a speech sample. The toolkit is accompanied by a number of trained models[5] of pitch accent and phrase boundary detection (and classification using ToBI labels); we use models for pitch accent detection and intonational phrase boundary detection, trained on three spontaneous speech corpora of Standard American English. The classifications are performed using the logistic regression algorithm with a range of pitch, intensity and duration input features [19]. The trained models were evaluated on all ten Hindi excerpts to derive labels indicating pitch accents and prosodic phrase boundaries.

4.2. Experimental results and discussion

Fig. 4 shows the kappa values for the automatically derived labels against the expert for both prominences and phrase boundaries; a confusion matrix with details of the insertion and deletion errors of AuToBI relative to the expert are also shown. We see that AuToBI almost never (only for 2 words) predicts a boundary when the expert does not. However, there are many instances (126 words) where AuToBI does not predict a boundary after the word while the expert does. These errors mainly stem from instances where a new prosodic phrase begins even when there is no preceding silence; this silence is an important feature for AuToBI to detect a boundary. For prominences,

AuToBI\Expert	Accent	No accent
Accent	74% (130/175)	37% (156/417)
No accent	26% (45/175)	63% (261/417)

Kappa coefficient: **0.311**

AuToBI\Expert	Boundary	No boundary
Boundary	43% (95/221)	0.5% (2/371)
No boundary	57% (126/221)	99.5% (369/371)

Kappa coefficient: **0.479**

Figure 4: Confusion matrix of AuToBI predictions against expert predictions, along with the kappa agreements, for both prominences and phrase boundaries.

[4] The speech files were sorted in descending order according to the Fleiss' coefficients of the "with audio" case.

[5] The toolkit and trained models are available at the following website: `http://eniac.cs.qc.cuny.edu/andrew/autobi/`.

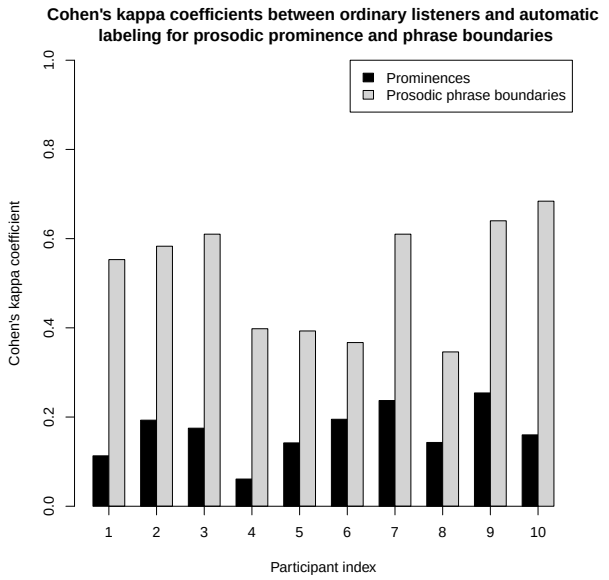

Figure 5: Cohen's kappa agreement coefficients for each participant against the automatically derived transcriptions for prosodic prominence and prosodic phrase boundaries.

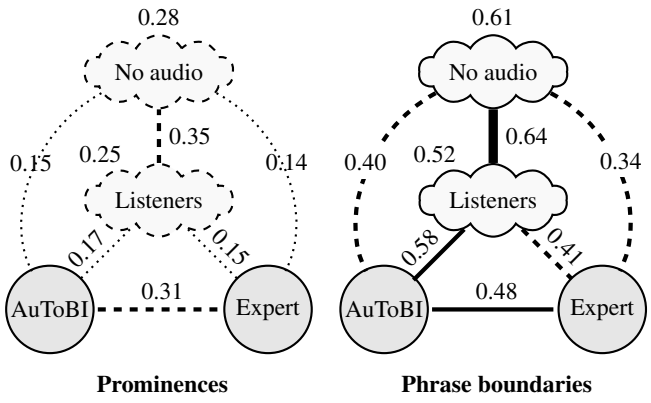

Figure 6: Kappa agreements between the non-expert transcribers, both using audio (*Listeners*) and without audio (*No audio*), AuToBI and the expert, for prosodic prominence and boundaries (shown on the left and right, respectively). The dotted lines indicate no agreement, the dashed lines indicate fair agreement, the bold lines indicate moderate agreement and the thick bold line indicates substantial agreement (according to the interpretation of the kappa statistic in [27]).

the false-positives result from words with a rising pitch accent which get classified as being prominent due to the pitch excursion (but are actually not prominent according to the expert).

Fig. 5 shows Cohen's kappa coefficients for each participant against the AuToBI predictions. As observed in Fig. 2, the listeners show a much higher value of agreement for phrase boundaries (mean $\kappa = 0.582$) than for prominences (mean $\kappa = 0.167$). Fig. 6 summarizes the agreement statistics between the non-expert transcriptions (both with and without audio), the expert transcriptions and the automatically derived transcriptions. We emphasize the following points:

1. The automatically derived labels for both prosodic events show fair to moderate agreement with the expert. This suggests the possibility of using AuToBI in the future for automatic prominence and boundary labeling in Hindi.

2. AuToBI predicts the non-expert transcribers' boundary scores better, but for prominence it is a better prediction of the expert's labels. This reaffirms the claim that ordinary Hindi listeners (unlike experts and machines) do not have a consistent internal definition of prominence.

3. The listeners are more in agreement with each other than with the expert; κ between the listeners and the expert for prominences fall within $[0.07, 0.24]$ while κ of the listeners with each other is in $[0.04, 0.49]$. This suggests that both are possibly tapping into different criteria for prosody perception.

4. In perceiving prosodic boundaries, the *Listeners* and *No audio* groups show moderate and substantial agreement with each other ($\kappa = 0.524$ and $\kappa = 0.612$, respectively). Further, for each participant, there is substantial agreement between the boundaries perceived with and without audio (κ is in the range $[0.55, 0.86]$). This suggests a fairly consistent bias amongst the listeners regarding what is expected of the task.

5. Listeners have much lower agreement for prominence than for boundaries. But the findings also show that, relative to the non-expert listeners, there is higher agreement between the expert transcriber and AuToBI on prominence (κ between listeners and AuToBI fall in the range of $[0.06, 0.25]$ showing that none of the listeners agree with AuToBI as much as the expert does, with $\kappa = 0.31$). This suggests that there are acoustic patterns in Hindi speech that are similar to the acoustic patterns that mark prominence in English, and further, that a trained Hindi speaker can discriminate among words on the basis of these acoustic patterns, as a basis for identifying prominence.

5. Conclusions and future work

We observe that non-expert listeners have much lower agreement for prominence than for boundaries amongst themselves as well as with the expert. On the other hand, AuToBI is more in agreement with the expert on prominence, relative to the non-experts. The fact that non-expert listeners fail to identify prominence on the basis of the same cues used by the expert and the machine suggests that either the patterns of acoustic prominence do not function to mark important linguistic information in Hindi, or they may serve multiple functions that are not easily lumped together in a single percept. Future research on Hindi is needed to investigate prominence under a wider range of pragmatic conditions (beyond contrastive focus), in production and perception.

We have found that automatic models of prosody for English make fairly good predictions about prosody in Hindi. We hope to improve on these models by fine-tuning them using labeled Hindi data; this would allow us to use relatively limited amounts of labeled Hindi data as opposed to building models of Hindi from scratch. We also propose to make use of these models in automatic speech recognition systems for Hindi.

6. Acknowledgements

This research was supported in part by a Beckman Postdoctoral Fellowship for the first author. The second and third author's contributions were supported by NSF BCS 12-51343 and QNRF NPRP 09-410-1-069, respectively. The authors gratefully acknowledge Tim Mahrt at the University of Illinois, Urbana-Champaign for developing the web software used in our perception study, Language Markup and Experimental Design Software (LMEDS).

7. References

[1] P. Moore, "A study of Hindi intonation," Ph.D. dissertation, University of Michigan, 1965.

[2] M. Ohala, "A search for the phonetic correlates of Hindi stress," C. M. Bh. Krishnamurti and A. Sinha, Eds., 1986, pp. 81–92.

[3] J. D. Harnsberger, "Towards an intonational phonology of Hindi," *Ms., University of Florida*, 1994.

[4] R. Nair, "Acoustic correlates of lexical stress in Hindi," in *Linguistic Structure and Language Dynamics in South Asia–papers from the proceedings of SALA XVIII roundtable*, 2001.

[5] L. O. Dyrud, "Hindi-Urdu: Stress accent or non-stress accent?" Ph.D. dissertation, University of North Dakota, 2001.

[6] U. Patil, G. Kentner, A. Gollrad, F. Kügler, C. Féry, and S. Vasishth, "Focus, word order and intonation in Hindi," *Journal of South Asian Linguistics*, vol. 1, pp. 53–70, 2008.

[7] V. Puri, "Intonation in Indian English and Hindi late and simultaneous bilinguals," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2013.

[8] S. Genzel and F. Kügler, "The prosodic expression of contrast in Hindi," in *Proceedings of Speech Prosody*, 2010.

[9] C. Féry, "Indian languages as intonational 'phrase languages'," in *Festschrift to honour Ramakant Agnihotri*, I. Hasnain and S. Chaudhury, Eds. Aakar Publisher, 2010.

[10] A. Sengar and R. Mannell, "A preliminary study of Hindi intonation," in *Proceedings of SST*, 2012.

[11] C. Féry and G. Kentner, "The prosody of embedded coordinations in German and Hindi," in *Proceedings of Speech Prosody*, 2010.

[12] J. Cole, Y. Mo, and S. Baek, "The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech," *Language and Cognitive Processes*, vol. 25, no. 7-9, pp. 1141–1177, 2010.

[13] M. Q. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," *Computer Speech & Language*, vol. 6, no. 2, pp. 175–196, 1992.

[14] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 2, no. 4, pp. 469–481, 1994.

[15] X. Sun, "Pitch accent prediction using ensemble machine learning," in *Proceedings of Interspeech*, 2002.

[16] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in *Proceedings of ICASSP*, 2004.

[17] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 216–228, 2008.

[18] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "TOBI: a standard for labeling english prosody," in *Proceedings of ICSLP*, 1992.

[19] A. Rosenberg, "AuToBI-a tool for automatic ToBI annotation." in *Proceedings of Interspeech*, 2010.

[20] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multilanguage telephone speech corpus," in *Proceedings of ICSLP*, 1992.

[21] J. Cole, T. Mahrt, and J. I. Hualde, "Listening for sound, listening for meaning: Task effects on prosodic transcription," To appear in *Proceedings of Speech Prosody*, 2014. [Online]. Available: http://prosody.beckman.illinois.edu/lmeds.html

[22] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," Version 5.3.51, retrieved from http://www.praat.org/.

[23] J. Cohen *et al.*, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[24] S. Calhoun, "Information structure and the prosodic structure of English: A probabilistic relationship," Ph.D. dissertation, The University of Edinburgh, 2007.

[25] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, vol. 1, no. 2, pp. 425–452, 2010.

[26] Y. Mo, J. Cole, and E.-K. Lee, "Naive listeners' prominence and boundary perception," *Proceedings of Speech Prosody*, 2008.

[27] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.