What Prosody Tells Infants To Believe

Elizabeth S. Kim
Department of Computer Science
Yale University
New Haven, CT 06520
Email: eliskim@cs.yale.edu

Kevin Gold
Department of Computer Science
Yale University
New Haven, CT 06520
Email: kevin.gold@yale.edu

Brian Scassellati
Department of Computer Science
Yale University
New Haven, CT 06520
Email: scaz@cs.yale.edu

***Abstract*—We examined whether evidence for prosodic signals about shared belief can be quantitatively found within the acoustic signal of infant-directed speech. Two transcripts of infant-directed speech for infants aged 1;4 and 1;6 were labeled with distinct speaker intents to modify shared beliefs, based on Pierrehumbert and Hirschberg's theory of the meaning of prosody [1]. Acoustic predictions were made from intent labels first within a simple single-tone model that reflected only whether the speaker intended to add a word's information to the discourse (high tone, H*) or not (low tone, L*). We also predicted pitch within a more complicated five-category model that added intents to suggest a word as one of several possible alternatives (L*+H), a contrasting alternative (L+H*), or something about which the listener should make an inference (H*+L). The acoustic signal was then manually segmented and automatically classified based solely on whether the pitches at the beginning, end, and peak intensity points of stressed syllables in salient words, were closer to the utterance's pitch minimum or maximum on a log scale. Evidence supporting our intent-based pitch predictions was found for L*, H*, and L*+H accents, but not for L+H* or H*+L. No evidence was found to support the hypothesis that infant-directed speech simplifies two-tone into single-tone pitch accents.**

I. Introduction

Prosody, or the melody of an utterance, can contain information about what the speaker thinks the listener should believe or know about an utterance. For example, when introducing herself for the first time, a speaker might say "Hello, I'm Eli Kim" in a high pitch, indicating that she believes this to be novel information to the listener. When giving a talk before an audience that already knows the speaker, however, the speaker might begin with a desultory "I'm Eli Kim" with low instead of high pitches on the name to indicate that the speaker expects the audience to know this. The present study is an exploration of whether similar signals about mutual belief exist in infant-directed prosody, and if so, whether such acoustic signals are simplified from their adult-directed versions. This study also describes a method by which to automatically detect mutual belief cues acoustically.

The literature on adult-directed prosody has produced a rich classification scheme to associate acoustic cues in speech with specific intents to modify the listener's and speaker's shared beliefs [1]. Meanwhile the literature on infant-directed speech has noted exaggerated prosodic features–a phenomenon known as "Motherese" [2]–inspiring explorations into turn-taking signals [3], speech stream segmentation [4], [5], signals to attract and maintain an infant's attention and communication of affect. There has also been investigation of infant-directed prosody as signals of new versus old information [6]. However, there has not been any investigation of whether infant-directed prosody contains the same variety of signals as does adult-directed speech, for intent to modify shared beliefs. Does infant-directed speech contain the same signals about mutual belief, or do parents simplify infant-directed prosody by reducing their selection of prosodic signals? The experiment described in this paper examines the prosodic patterns of infant-directed speech taken from the CHILDES database [7] in order to determine whether infants as young as 16 months receive the full variety of pitch accents that signal, in adult-directed speech, speaker intent to modify shared knowledge.

This work should be of interest to developmental psychologists and developmental roboticists alike, as prosodic cues about a speaker's expectations of the listener's knowledge could be a powerful learning aid to infants and infant-like learners. When a speaker teaches an infant new words or new facts, prosody could help the learner determine whether the item to be learned is novel, or is a variation or new example of an old fact. This, in turn, could help reduce learner errors; if the speaker mishears a word as "abul" but the prosody indicates that the item should be familiar, the learner may be able to interpret the misheard word as a familiar word, "apple." Such aids to learning are overlooked in approaches to prosody that account only for information about affect or mood only, as is common in robotic learners [8], [9], [10], [11].

Early exposure to prosodic cues about shared belief could also help infants develop an ability to reason about beliefs, sometimes called "theory of mind" [12]. Knowing whether such prosodic cues exist in infant-directed speech is critical for models of infant development of theory of mind. If infant-directed prosody contains the same mutual belief cues as does adult-directed prosody, this may be a key source of input to a learner that is developing a theory of mind.

Section II will provide some background about prosody, including the labelling scheme of Pierrehumbert and Hirschberg [1] which we will use extensively in the following sections. Our semi-automatic acoustic classification method will also be described in Section II. Section III will describe our experiment in which audio data from the CHILDES corpus was analyzed to determine whether the infant-directed speech matched the predictions implied by Pierrehumbert and Hirschberg's system. Section IV will include our analysis of

the data, and our conclusions in Section V will return to the question of whether infant-directed speech contains the same cues to mutual belief as adult-directed speech.

II. Pitch Accents, Prosody, and Cues for Shared Knowledge

Prosody is the music of speech. It is manifested in variations of pitch, loudness, duration of syllables and pauses, and voice quality. In English prosody is somewhat determined by linguistic considerations, such as stress on syllables within words, and question versus non-question information. Otherwise, English prosody flexibly conveys paralinguistic or nonlinguistic information, such as the speaker's intention or attitude, and mood or affective state [13].

Pitch is the highness or lowness of the voice, sometimes called the tune or melody of an utterance. Pitch is a percept which roughly correlates acoustically with the fundamental resonant frequency $f0$ of voiced phonemes, including vowels, nasals (/m/ and /n/), voiced obstruents (e.g., /b/) and approximants (e.g., /l/). Most adult male voices vary in pitch over frequencies from 50 to 300 Hz. The pitch of adult females and children can range from 150 to 1000 Hz [14].

A. Previous Research in Infant-Directed Prosody

In the robotics and cognitive science literature, previous computational research on infant- or infant-like-learner-directed speech has largely focused on communication of mood or affective intent. Systems have been built to recognize or describe the prosody of speaker approval (with sustained pitch peak intensity) and prohibition (with low, staccato tones) [15], [9], [8], [11], [10], bids to attract attention (with rising pitch contours) [9], [8], [11], [16], and soothing intent (with falling pitch contours) [8], [11], [17]. Infant-directed prosody has also been investigated for cues to turn-taking [3], speech stream segmentation [4], [5], and new versus old information [6].

There has been limited investigation into shared belief cues in infant-directed prosody. Adult-directed prosody is thought to convey information about what is mutually believed between speaker and listener (see Section II-B). Whether such signals exist in infant-directed prosody has not been previously studied in the framework we discuss below, but there might be good reason to think that adults might modify their prosody to make it less complex. It is known that parents tend to speak to their children in exaggerated prosody, known as "Motherese" [2]. Compared with adult-directed prosody, infant-directed prosody features higher pitch and wider pitch range [4], [18], [19], [20], and longer vowels at phrase [21] and clause boundaries [22]. This prosodic exaggeration has inspired some builders of robotic prosody classifiers to attempt to elicit Motherese-like speech with infant-like robots[8], [10].

Whereas investigations of Motherese have suggested that infant-directed pitch contours are characteristic of specific affective intents, an alternative view may be that these contours are determined by the informational content of the speech. Pierrehumbert and Hirschberg have argued that the tune of adult-directed prosody cannot be explained either in terms of the speaker's speech acts or emotion, since the mapping from tune to speech act or emotion is at best one-to-many [1]. Instead, to describe adult-directed prosody they proposed the system described below, in which prosody signals each word's relation to the speaker's intended modification of shared beliefs. To our knowledge, the interaction of the system described in [1] with infant-directed effects has not been explicitly studied before, though some similar observations about novelty affecting pitch have been made in the infant-directed literature [6].

B. Prosody, Shared Beliefs, and Discourse Structure

The following exposition of the shared belief information in prosody is based closely on that of Pierrehumbert and Hirschberg [1], which has been empirically supported to some extent [23]. The labeling scheme summarized here is the basis for the popular ToBI representation of prosody [24], [25].

In English a speaker produces a *pitch accent* for at least one word in each utterance, marking it as salient. A high or low pitch on the stressed word conveys whether the speaker intends for the listener to add the word's information to their *mutual beliefs* [1]. Accented words are perceived by listeners to be *prominent*, or *stressed*, with relation to other words. In English, every word has at least one stressed syllalble; however, accented words receive an additional stress over other words. Stress of one word over others is conveyed though a combination of greater loudness, longer duration, and hyperarticulation of that word's stressed syllable. There are two simple pitch accents, H and L, and three two-tone pitch accents, which combine H and L pitches.

The H* pitch accent is used to convey the speaker's intent for the listener to add the accented information to their shared beliefs. Perceptually, an H*-accented word will feature a relatively high pitch at the perceptually prominent syllable in the prominent word. The '*' diacritic indicates temporal alignment with the stressed syllable. This accent is commonly used when introducing new information, and frequently appears in declarative statements. For instance,

Alice	likes	Bob
H*		H* L L%

Here, the speaker S intends for the listener L to add the fact of Alice's liking and the fact that Bob is liked to L's beliefs. This utterance thus would be appropriate if, for example, neither person had been mentioned in the conversation previously. (The L L% at the end refers to the pitch of the phrase and whole utterance, respectively; we include these markings for completeness but will not discuss them.)

H* can be used to add connoted rather than denoted information to L's beliefs. For instance, in this example, S tells L what L has done (and thus presumably already knows). Here S uses an H* accent to convey that L should add knowledge of S's awareness to L's beliefs.

You	ate	my	cookie on	purpose
H*	H*		H*	H* L L%

The L* simple pitch accent is perceptually indicated by a prominent word that is close to the baseline pitch for the speaker. It indicates the speaker's intent for the listener not to add the accented item to his beliefs. This accent is commonly used when S is uncertain, such as in yes or no questions:

Did	our	paper	get	rejected
		L*		L* H H%

L* can indicate S's belief that the expression is incorrect:

I	guess	our	paper	just	isn't	good	enough
	L*		L*		L*	L*	L* L L%

or when uttering information believed already known by L:

I'd	like	coffee	and I think I'll have a muffin
L*	L*	L* H H%	

In all these cases, S does not intend for L to add the L* accented information to their shared beliefs, since the L* accented items are either uncertain, false, or previously added.

In two-tone, as in single-tone, pitch accents the '*' indicates temporal alignment with the stressed syllable. The L*+H pitch accent perceptually is perceived as a low frequency on a stressed syllable, followed immediately by a rising pitch contour to a higher pitch. L*+H pitch accents indicate uncertainty in an implied comparison of scale. For example,

A:	This	talk	is	terrible.
B:	The	paper	was	good
				L*+H L H%

The L*+H accent on good indicates S's uncertainty as to the relevance of the paper's quality to the quality of the talk.

Likewise, L+H* pitch accents also signal an intended comparison of scale, but are instead conveyed with certainty, expecting the listener to add the accented item to S and L's shared beliefs. For example,

A: This	paper	is	awfully	informal.	
B: It's	even	chatty	for a	conference	paper
				L+H*	L L%

The H*+L accent signals that the listener should infer support for the accented items from previously existing beliefs. Like the H* accent, H*+L signals that S should add the accented item to their beliefs, but also should make an inference based on the new information and existing beliefs, such as an implied course of action:

Your	dinner's	getting	cold
	H*+L	L*	H*+L H L%

Pierrehumbert and Hirschberg suggested that H+L* possessed a similar meaning to H*+L, but was used to convey information already known to the speaker [1]. However, [1] also noted that "there is some difficulty in separating the meaning of H+L* from that of H*+L, because in many cases the phonological analysis is unclear" (p. 300). Moreover, in modern labeling conventions, the H+L* notation has been superseded with H+!H*, to note that this contour usually remains higher than other low tones [26]. For these reasons, this tone was not predicted for any utterances in our experiment, though we did check for acoustic evidence of it.

C. Acoustic Methodology

Usually pitch accents are manually classified by trained specialists using acoustic recordings and graphical representations of the pitch contour over time [25]. Manual classification tends not to produce high amounts of agreement among experts [27].

Attempts to automate pitch accent labeling from acoustic features have tended to focus on locating, rather than classifying, pitch accents [28], [29], have classified only a very limited subset of pitch accents for adult-directed speech [30], or have classified pitch accents for languages beside English [31]. Pitch accents have been statistically clustered with high agreement (78%) with listeners' judgments, suggesting acoustic regularities distinguishing pitch accent categories [32].

For this paper, we introduce a partially automated method that allows American English pitch accents to be acoustically, quantitatively classified. This allows our hypothesis-testing to be free of bias introduced by our knowledge of the semantic content of the speech, and also is a step toward the fully automated classification of pitch accents. (The reader may find it useful to refer to Figure 1.)

To begin, intensity and $f0$ contours over time are estimated using Praat phonetic software [33]. Utterances in the CHILDES transcripts are linked to their temporal positions in the recordings. CHILDES' CLAN software links text to Praat.

Next, utterance-minimum and -maximum $f0$ are extracted, using Praat, giving a baseline pitch and pitch range for the speaker at the time of utterance. Measuring pitch range locally within each utterance, instead of over the speaker's entire history, allows the range to adapt to the speaker's current affect and immediate auditory conditions, though it has the disadvantage of sometimes producing an falsely small range for utterances having no H* pitches. During this stage utterances within which $f0$ estimation software clearly fails, are manually discarded: failures include sudden pitch drops below 75 Hz, sudden jumps to overtones (doubling or halving errors), or misclassification of unvoiced noise as pitch. (Section IV describes how frequently this occurred in our experiment.)

The next step is the segmentation of the stressed syllables in the selected words. Segmentation was done manually by listening to the audio and using cues from the intensity curves (e.g., "stops" such as "p" and "k" literally stop the air momentarily, and thus are clearly marked by low intensity). The stressed syllable of a selected word is the relevant part of the audio signal for pitch accent classification.

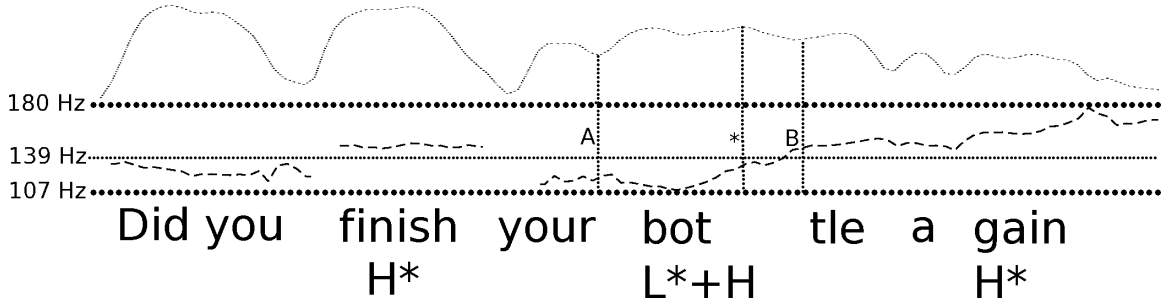

Fig. 1. A sample utterance from MacWhinney's CHILDES corpus, with intensity (top) and $f0$ (middle, dashed) extracted using Praat phonetic analysis software. The minimum and maximum $f0$ of the utterance establish the baseline and range, and their average on a log scale gives the dividing line between L* and H* pitches. For two-tone classifications, the $f0$ at the beginning of the stressed syllable (A) gives the first tone, and the end of the syllable (B) gives the second; the stress is placed based on the syllable's point of maximum intensity (*). Though the statement is phrased as a question (suggesting L*), in fact the speaker is essentially telling the infant that he is aware that the infant is done (H*) but is unsure whether the whole bottle is gone (L*+H).

To classify a pitch accent as a single tone, $f0$ at the point of maximum audio intensity within the syllable is compared to the baseline minimum and maximum $f0$ over the whole utterance. After taking the logarithm of all three fundamental frequencies–minimum, maximum, and $f0$ at time of maximum intensity–whether the maximum-intensity-$f0$ is closer to the baseline minimum or maximum determines whether it is L* or H*. This comparison is done on a log scale, a method we introduce here for pitch accent classification, because just-noticeable-differences for pitch are logarithmic in frequency in the 50-5000 Hz range [34], which covers the range of human speech, and because H* pitch accents are thought to actually be medium to high pitches within the speaker's range [26], which intuitively fits well with a log-scale model.

To classify a pitch accent as two-tone, our method examines $f0$ at the beginning and end of the syllable as well. These points, which were manually identified for syllable identification, are subjected to the same logarithmic transformation, and classified as L or H based on whether they are above or below the log-transformed midpoint of the speaker's range. If the two endpoint classifications are the same, the pitch remains classified as a simple L* or H*. If they are different, then the pitch is classified as a two tone accent, L+H or H+L. In the L+H case, the location of the accent mark is determined by the classification of the maximum intensity point. If the pitch at the time of maximum intensity is closer to the log-transformed pitch baseline, it is L*+H; otherwise, it is L+H*. The maximum intensity classification is similarly used to distinguish between H*+L and H+L*.

Both the acoustically simple one-tone method and the two-tone method were used and compared to our theoretical predictions in the experiments to be described below.

III. Experiment

Two transcripts of infant-directed speech from the CHILDES database [7] were examined: one of a father speaking to his 16-month-old son [35] and another of a mother speaking to her 18 month-old-daughter [36]. 165 words from these transcripts were chosen as targets for comparison of the theoretical predictions of the Pierrehumbert and Hirschberg model [1] to the observed acoustics. A word was chosen as a prediction target if it was central to the meaning of its sentence, and if the transcript context made one pitch accent category seem more likely than the others. Single- and two-tone predictions were made for each selected word, given the conversational situation. In the forced two-choice prediction, pitch accent was predicted depending on whether the text suggested that the parent wished to introduce informational content with the word or not. In the five-category prediction, the experimenters made their predictions based on whether the word was being tentatively suggested as one of several specific alternatives (L*+H), being specifically suggested in contrast to another alternative (L + H*), was a reminder of something that the child should already know (H+L), or was otherwise introducing new information (H*) or not (L*). Predictions were made based on the textual transcripts alone, without having heard the audio recordings.

We note that our predictions assumed neither an accurate representation of the infant's belief state on the part of the speaker, nor expectations on the part of the speaker of adult-like belief state for the infant listener. Rather, we assumed that speakers tailor their representations of the listener's belief states to the individual listeners and conversations. Our predictions reflect only indications from local context in the transcripts (of up to a few preceding and following sentences) about the speaker's intents to modify what they apparently conceived to be mutual beliefs.

We also distinguish between introduction of a new word (for example, the naming of a novel object) and new information, a broader act, which can include, for example, newly achieved certainty in interpreting an infant's proto-linguistic requests for a bottle. Our H* predictions are of the broader sort.

Following transcript-based predictions, the utterances containing the selected words were then analyzed using the acoustic method introduced in section II-C, and the quantitative results compared to the theoretical predictions.

IV. Results

Of the 165 utterances, 29 utterances were discarded because of audio noise or incorrect segmentation within the corpus,

leaving 136 data points for each of the single-tone and two-tone classification schemes.

The single-tone predictions of L* and H* coincided with the results of our single-tone acoustic analysis method (see Section II-C) for 87 of the words, or 64% of the time; this was significantly more often than chance ($\chi^2 = 8.61, p < 0.005$). Though we had entertained the hypothesis that the difference might be attributed to whether the word was contained within a question or not, there was no evidence to support this idea ($\chi^2 = 1.46$, $p = 0.228$).

56 of the 136 two-tone predictions were correct, an occurrence highly unlikely to be due to chance because of the five categories ($\chi^2 = 57, df = 20, p < 0.001$). Broken down into category by category comparisons, we found that the H*, L*, and L*+H predictions each produced significantly more correct responses than could be attributed to chance ($p < 0.005, p < 0.001, p < 0.005$, respectively), while the L+H* and H*+L predictions provided no such evidence of accuracy ($p = 0.656, p = 0.561$).

However, there was no evidence to support the hypothesis that parents tended to simplify their pitch accents toward their children, as there was no evidence that single tones were more likely to be observed in the place of two-tone accents than vice versa ($\chi^2 = 0.459, p = 0.498$).

Qualitatively, H* and H*+L were common pitch accents for introducing or reinforcing labels for objects (annotations are those provided by our acoustic method):

CHILD:	What's	that?		
FATHER:	Tape	recorder	over	there
	H*	H*+L		

L* was most common in cases when the parent was offering an interpretation of what the child was communicating:

FATHER:	You	like	the	soldiers?
		L*		L*

However, L* also occurred where we had predicted H* in cases where it seemed from the text that the parent was pointing out new information, but the parent was actually going through a ritual such as reading a familiar book:

MOTHER:	And that's a	rabbit	with no	face.
		L*		L*

L*+H was often used in its adult meaning of an alternative that the speaker was unwilling to support, but in cases where one might have expected L+H* to indicate correction, the speaker did not appear to follow through:

CHILD:	dog? ...		
FATHER:	is that a	doggy	Honey ? ...
		L*+H	
FATHER:	or is that	[//] he's a	kitty ?
			L*

TABLE I
INCIDENCE OF PREDICTIONS AND OBSERVATIONS FOR PIERREHUMBERT AND HIRSCHBERG'S SIX CATEGORIES OF PITCH ACCENT.

Pred \ Obs	H	L	L*+H	L+H*	H*+L	H+L*	Total
H	15	8	0	2	3	0	28
L	10	30	4	4	4	2	54
L*+H	8	6	7	2	0	2	25
L+H*	3	1	3	2	6	0	15
H*+L	5	3	1	0	2	2	13
H+L*	0	0	0	0	0	0	0
Total	41	48	15	10	15	6	135

H*+L was very occasionally observed in the role of asking the child to make an inference, but this was not consistent:

MOTHER [pointing to	mirror]: Who's in	there? ...
		H*+L
MOTHER: That's	Amelia!	
	H*	

As these examples illustrate, the instances in which the predictions failed to match the observations were often explainable by the ambiguity of the text, and not a failure of the theory or acoustic method.

V. CONCLUSIONS

These results demonstrate that at least some of Pierrehumbert and Hirschberg's acoustic signals about shared belief, and our acoustic method for identifying pitch accents, hold for American English infant-directed prosody at ages 16-18 months. Our data shows strong evidence for H*, L*, and L*+H accents' usage for conveying the same information about mutual belief proposed in the adult-directed case, at least for the two speakers whose prosody we investigated thoroughly. These differences in pitch are not determined by a word's embedding in a question, but mark whether or not the speaker wishes to introduce new information with the word, or (in the case of L*+H) whether the speaker offers the word as one of several possible alternatives.

Our findings show that even when speaking to infant listeners, with immature cognitive and linguistic capabilities, speakers signal their intent to modify listener's beliefs, in ways similar to those suspected to be used for adult listeners. In other words, infants are receiving cues about what is shared information even at an age when they are unlikely to have a concept of distinct states of knowledge between distinct individuals, which is demonstrated considerably later [12]. It is therefore possible that children use pitch accent signals in learning to reason about shared and private information. Understanding the role of prosody in this process of reasoning about shared knowledge may be critical to understanding how "theory of mind" develops, and also to understanding why and how autistic children tend to demonstrate an impaired ability to reason about minds. A better understanding of how typical children integrate and react to infant-directed prosody may

help early diagnosis of autism, which is known to include abnormal prosody as one of its symptoms [37].

The lack of evidence for H*+L and H+L* supports a recent tendency to view these categories of [1] as less well supported by the data than the other pitch accents [38], but it is somewhat unclear why L+H* poorly matched our predictions. There are several possible explanations. These accents may be particularly difficult to accurately predict from transcripts, since the difference between L*, L*+H, and L+H* might depend on how strongly the parent prefers an alternative. It is also possible that our acoustic method does not accurate describe L+H* accents. It is also possible that parents intentionally avoid this contour because of its negative connotation as a correction. This is a good question for future study.

What is clear is that American English infant-directed prosody contains some of the interesting signals about shared information theorized to exist in adult-directed prosody, and that a relatively simple method–comparing the log of the maximum intensity pitch to the speaker's maximum and minimum pitches–can extract them.

It may therefore be useful for creators of robotic systems to bear pitch accents in mind as an additional source of speech information. Though we have not yet measured agreement between our acoustic method with trained listeners' pitch accent judgments, our method offers quantitative, acoustic information about speaker intent. Automating the manual parts of our method, namely stressed syllable segmentation and discarding noise, are areas for future work.

ACKNOWLEDGMENTS

This work was supported by a National Science Foundation CAREER award (#0238334), NSF award #0534610 (Quantitative Measures of Social Response in Autism), a Microsoft Research Human-Robot Interaction award, a software grant from QNX Software Systems Ltd, and the Sloan Foundation.

REFERENCES

[1] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in interpretation of discourse," in *Intentions in Communication* (P. R. Cohen, J. Morgan, and M. E. Pollack, eds.), pp. 271–311, Cambridge, MA: MIT Press, 1990.

[2] A. Fernald, "Four-month-old infants prefer to listen to motherese," *Infant Behavior and Development*, vol. 8, pp. 181–195, 1985.

[3] C. E. Snow, "The development of conversation between mothers and babies," *J. Child Language*, vol. 4, pp. 1–22, 1977.

[4] A. Fernald and T. Simon, "Expanded intonation contours in mothers' speech to newborns," *Developmental Psychology*, 1984.

[5] E. D. Thiessen, E. Hill, and J. R. Saffran, "Infant-directed speech facilitates word segmentation," *Infancy*, vol. 7, no. 1, pp. 53–71, 2005.

[6] A. Fernald and C. Mazzie, "Prosody and focus in speech to infants and adults," *Developmental Psychology*, vol. 27, no. 2, pp. 209–221, 1991.

[7] B. MacWhinney, *The CHILDES project: Tools for analyzing talk*, vol. 2. Mahwah, NJ: Lawrence Erlbaum Associates, 3rd ed., 2000.

[8] C. Breazeal and L. Aryananda, "Recognizing affective intent in robot directed speech," *Autonomous Robots*, vol. 12, no. 1, pp. 83–104, 2002.

[9] M. Slaney and G. McRoberts, "Baby ears: A recognition system for affective vocalizations," *Speech Communication*, vol. 39, pp. 367–384, Feb 2003.

[10] E. S. Kim and B. Scassellati, "Learning to refine behavior using prosodic feedback," in *Proc. Int. Conf. Development and Learning (ICDL)*, 2007.

[11] A. Robinson-Mosher and B. Scassellati, "Prosody recognition in male infant-directed speech," in *Proc. the 2004 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2004.

[12] A. Gopnik, "Theory of mind," in *The MIT Encyclopedia of the Cognitive Sciences (MITECS)* (R. A. Wilson and F. C. Keil, eds.), Cambridge, MA: MIT Press, 2001.

[13] H. Mixdorf, "Speech technology, tobi and making sense of prosody," in *Speech Prosody*, vol. 3, pp. 31–38, 2002.

[14] T. F. Quatieri, *Discrete-Time Speech Signal Processing, Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 2002.

[15] D. Roy and A. Pentland, "Automatic spoken affect analysis and classification," in *Int. Conf. Auto. Face Gesture Recognition*, pp. 363–367, 1996.

[16] L. J. Ferrier, "Intonation in discourse: Talk between 12-month-olds and their mothers," in *Children's Language* (K. Nelson, ed.), vol. 5, pp. 35–60, Hillsdale, NJ: Erlbaum, 1985.

[17] M. Papousek, H. Papousek, and M. H. Bornstein, "The naturalistic vocal environment of young infants: On the significance of homogeneity and variability in parental speech," in *Social Perception in Infants* (T. Field and N. Fox, eds.), vol. 5, pp. 269–297, Norwood, NJ: Ablex, 1985.

[18] O. Garnica, "Some prosodic and paralinguistic features of speech to young children," in *Talking to children: Language input and acquisition* (C. E. Snow and C. A. Ferguson, eds.), Cambridge, England: Cambridge University Press, 1977.

[19] L. Menn and S. Boyce, "Fundamental frequency and discourse structure," *Language and Speech*, vol. 25, pp. 341–383, 1982.

[20] D. M. Stern, S. Spieker, R. K. Barnett, and K. MacKain, "The prosody of maternal speech: Infant age and context related changes," *J. Child Language*, vol. 10, pp. 1–15, 1983.

[21] J. L. Morgan, *From simple input to complex grammar*. Cambridge, MA: MIT Press, 1986.

[22] N. Bernstein Ratner, "Durational cues which mark clause boundaries in mother-child speech," *J. Phonetics*, vol. 14, pp. 303–309, 1986.

[23] M. S. Emiel Krahmer, "On the alleged existence of contrastive accents," *Speech Communication*, vol. 34, pp. 391–405, Jul 2001.

[24] K. Silverman, M. E. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "TOBI: A standard for labeling english prosody," in *Int. Conf. Spoken Language Process. (ICSLP)*, vol. 3, pp. 235–238, 1992.

[25] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The original ToBI system and the evolution of the ToBI framework," in *Prosodic Typology: The Phonology of Intonation and Phrasing* (S.-A. Jun, ed.), ch. 2, Oxford, England: Oxford University Press, 2005.

[26] J. Hirschberg and M. E. Beckman, "ToBI annotation conventions." 1994.

[27] A. K. Syrdal and J. McGory, "Inter-transcriber reliability of ToBI prosodic labeling," in *Int. Conf. Spoken Language Process. (ICSLP)*, vol. 3, pp. 235–238, 2000.

[28] S. Ananthakrishnan and S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, pp. 216–28, Jan 2008.

[29] M. Hasegawa-Johnson, K. Chen, J. Cole, S. Borys, S.-S. Kim, A. Cohen, T. Zhang, J.-Y. Choi, H. Kim, T. Yoon, and S. Chavarria, "Simultaneous recognition of words and prosody in the boston university radio speech corpus," *Speech Communication*, vol. 46, pp. 418–439, 2005.

[30] X. Sun, "Pitch accent predictions using ensemble machine learning," in *Int. Conf. Spoken Language Process. (ICSLP)*, 2002.

[31] B. Kim and G. G. Lee, "C-tobi-based pitch accent prediction using maximum-entropy model," in *Lecture Notes Comp. Sci. (ICCSA)*, vol. 3982, (Heidelberg, Germany), pp. 21–30, Springer Berlin, 2006.

[32] G.-A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent," in *Proc. Human Language Technology Conf. N. Amer. Ch. Assoc. Comp. Linguistics*, (Morristown, NJ), pp. 224–231, 2006.

[33] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (v. 5.0.01)," 2007.

[34] G. M. Clark, *Cochlear Implants: Fundamentals and Applications*. New York: Springer, 2003.

[35] B. MacWhinney, "Childes database manual: American english." 1994.

[36] N. Bernstein Ratner, "The phonology of parent child speech," in *Children's Language* (K. Nelson and A. vanKleeck, eds.), vol. 6, 1987.

[37] L. D. Shriberg, R. Paul, J. L. McSweeny, A. Klin, D. J. Cohen, and F. R. Volkmar, "Speech and prosody characteristics of adolescents and adults with high-functioning autism and asperger syndrome," *J. Speech, Language, and Hearing Research*, vol. 44, pp. 1097–1115, Oct 2001.

[38] M. E. Beckman and G. A. Elam, "Guidelines for ToBI labelling, v.3." 1997.