Quantifying developmental changes of prosodic categories

Britta Lintfert[1], *Antje Schweitzer*[1], *Lukasz Wolski*[2], *Bernd Möbius*[1 2]

[1]Institute of Natural Language Processing, University of Stuttgart, Germany
[2]Department of Speech and Communication, University of Bonn, Germany
britta.lintfert@ims.uni-stuttgart.de, antje.schweitzer@ims.uni-stuttgart.de, lwolski@uni-bonn.de , moebius@ifk.uni-bonn.de

Abstract

The aim of this case study is to introduce a new automatic method for describing the development of prosodic categories in speech acquisition. We use the PaIntE model (Parametrized Intonation Events) to examine fine phonetic detail in one child's F0 contours at several stages between 7 and 22 months of age. The variability in these contours is quantified using K-means clustering. In contrast to traditional contour-based or autosegmental-metrical based descriptions of the development of intonation, this method can be applied to both babbling and more complex multi-word utterances, which is favorable for longitudinal studies of intonation in child speech.
Index Terms: speech acquisition, intonation, parametric approach, clustering methods

1. Introduction

In traditional contour-based descriptions, the development of intonation is described in terms of rise and fall of F0 ("contour shape") and their amplitudes in semitones ("contour range") at different developmental stages. The contour range can be compared to mature targets [1], or the contour inventories of the children at different stages are described independently [2].

This holistic approach is useful for describing contour shapes in babbling and one-word utterances, but it fails in more complex multi-word utterances because it does not make any assumptions about where the sentence accent is placed. To our knowledge, no studies have examined the systematic use of contour shapes depending on their discourse function. But as more and more words are produced, the function of intonation in discourse has to be taken into account.

For this reason more and more studies investigate applying the intonational categories of adult speech posited by the ToBI labeling system [3] to child speech [4, 5]. These categories are claimed to serve different higher-linguistic functions, for instance in the domain of discourse interpretation.

However, the categories described by ToBI or by its language-specific variants are developed for adult speakers. The problem in applying adult categories to child speech is the assumption that children with the beginning of meaningful speech are already capable of consistently using the different categories of intonation. Using these categories for child speech does not account for possible other categories during the acquisition of intonation based on children's limitations in production. Even the first child productions during babbling show language-specific relations between the prosodic structure and the communicative intentions of the child [6]. However, these productions are constrained by the motor abilities of the developing articulatory system.

Against this background, we introduce a new automatic method for describing the F0 contours during speech acquisition. We parametrized F0 contours of one boy between 7 and 22 months of age using the PaIntE approach [7]. We then tried to identify groups of similar contours using K-means clustering, claiming that different clusters may be interpreted as different intonational categories. These hypothetical categories were then compared to existing GToBI(S) categories based on the typical PaIntE parameters observed for these categories in adult speech [8].

GToBI(S) [9] is an adaptation of ToBI to German. GToBI(S) provides 5 basic types of pitch accents with different discourse interpretations: L*H, H*L, L*HL, HH*L, and H*M. These contours can also be described as rise, fall, rise-fall, early peak, and stylized contour, respectively. For L*H and H*L, allotonic variants exist, for instance, monotonal L* for L*H, or monotonal H* for H*L.

2. Method

2.1. Participant and data collection

This case study is based on longitudinal data of one typically-developed monolingual German boy (aged 0;7–1;10).The data are part of the Stuttgart Child Language Corpus [10]. The recordings took place at the boy's home in familiar play situations with his parents. The boy was recorded during interactions with his parents while looking at picture books or playing with toys. Thus the data represent spontaneous productions of the boy. However the setting was controlled to some degree because the parents were always using the same picture book during the babbling and first word production phase to motivate comparable productions from the child.

All recordings were transferred to a computer workstation, downsampled to 16 kHz and manually annotated on the segment, syllable and word level.

2.2. PaIntE parametrization

PaIntE stands for "Parametrized Intonation Events" [7] and was originally developed for F0 modeling in speech synthesis. PaIntE approximates stretches of F0 by a phonetically motivated function which is the sum of a rising and a falling sigmoid with a fixed time delay. The parametrization uses six parameters, viz. the height of the F0 peak (parameter d), the temporal position of the peak in the syllable (b), and the amplitudes (c1, c2) and the steepness (a1, a2) of the rising and falling sigmoid. A schematic of the function is given in Figure 1. The time axis

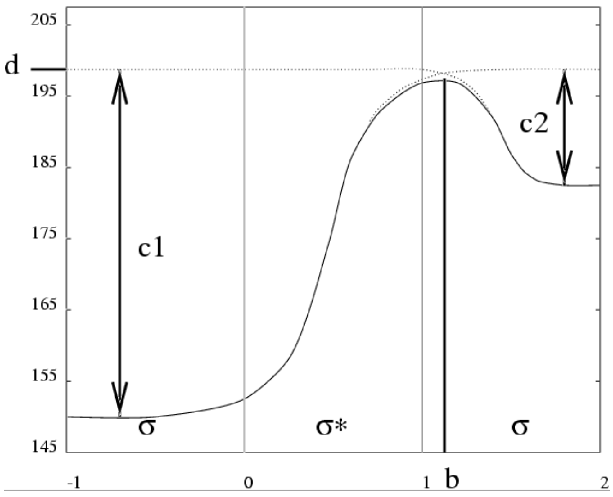

Figure 1: *Schematic of the PaIntE approximation function, reproduced from [7]. The approximation window represents three syllables, where the accented syllable is indicated by the asterisk (σ*). Peak height is determined by parameter d, amplitudes of rise and fall correspond to parameters c1 and c2, respectively, and peak alignment depends on the b parameter.*

is normalized to the lengths of syllables, i.e., the peak is at the beginning of the accented syllable if b=0, and at its end if b=1.

Based on a corpus of one adult speaker with approximately 6200 pitch accents [8] has shown that peak alignment (parameter b), peak height (d), and the amplitudes of rise (c1) and fall (c2) capture the tonal properties attributed to the different intonation events posited by GToBI(S). Also, different probability distributions of the PaIntE parameters were observed for different GToBI(S) pitch accents. For instance, as expected, for falling (H*L) accents, c2 is greater than c1, and the b parameter indicates that the peak is approximately in the middle of the accented syllable, while for rising (L*H) accents, c1 is greater than c2, and the b parameter is greater than in the H*L case, indicating that the peak occurs later than for H*L accents. Based on these observations for adult speech, we tried to assign GToBI(S) events to the clusters found in child speech.

As in [8] the parametrization was carried out for every syllable of the speech data, always using a three-syllable window. The method has never been applied to child language before; therefore, no comparable results are available.

2.3. Cluster analysis and interpretation of clusters

Using the kmeans function of R, we conducted K-means cluster analyses based on the parameters c1, c2 and b, since these parameters were useful in distinguishing the underlying intonational events for adult speech [8]. The cluster analysis was performed for each developmental stage separately on all syllables to separate different intonational categories for each stage. No initial cluster centers were specified. For each stage the analysis was performed with 18 iterations and a final change in cluster centers of zero. The optimal number of clusters was based on the homogeneity of variance of the clusters: we selected the minimal cluster size for which F<1 for each cluster.

Two properties were derived from the cluster centers. First, we compared the c1 and c2 parameters of the cluster centers to assess whether the contour corresponds to a rise-fall (both c1 and c2>30 Hz), or else to an overall fall (c2>c1), or an overall rise (c1>c2). Syllables for which both c1 and c2 were less than 20 Hz were interpreted as unaccented with no specific contour shape.

no.	size	b	c1	c2	shape	range	GToBI(S)
C1	2	0.41	112.39	91.56	c1,c2>30 ⇒ rise-fall	wide	L*HL
C2	7	0.62	67.17	0	c1>c2 ⇒ rise	wide	L*H
C3	11	0.24	10.77	42.97	c1<c2 ⇒ fall	narrow	H*L

Table 1: Clusters at **7 months** with cluster sizes, c1, c2, and b parameters of the cluster centers, contour shape, contour range, and corresponding GToBI(S) events

Second, the categorized contours were classified according to their range as *wide* (indicating high maturity) when the accent range was more than 4 semitones for falls and 3 semitones for rises. Contours with smaller ranges were classified as *narrow* (indicating adult-like contour shape but low maturity) [1]. For rise-fall shapes, we observed some cases where the range of fall and rise contradicted each other, one being narrow and the other being wide. In these cases, we classified the range as mid range (indicating mid maturity).

To transform the accent range before the peak (1) and after the peak (2) from Hertz to semitones, we used parameter d, which indicates the absolute F0 value of the peak in Hz.

$$range(c1) = \frac{12}{log(2)} * (log\frac{d}{d-c1}) \quad (1)$$

$$range(c2) = \frac{12}{log(2)} * (log\frac{d}{d-c2}) \quad (2)$$

3. Results

Tables 1 to 7 present the clusters for each stage. The c1, c2, and b parameters of the cluster centers in Hz are indicated in each table along with the overall characterization of the corresponding contour as rise, fall, or rise-fall based on the ratio of c1 and c2 ("shape") and accent range ("range") as described in Section 2.3. The final column indicates our interpretation in terms of GToBI(S) categories.

At the age of 7 months three different clusters can be found (Table 1). We interpret cluster C1 as instances of L*HL since c1 and c2 are similarly high and both greater than 30 Hz, which indicates a rise-fall contour based on the criterion formulated above. Further evidence for this interpretation comes from the fact that b = 0.41 indicates that the peak of the accent is approximately in the middle of the accented syllable, which was typical for L*HL in [8]'s data. For an L*H, in contrast, the peak was found to be close to the end of the accented syllable in word-final contexts or on the post-accented syllable in word-medial contexts. Also, c1 for L*H tended to be between 20 and 60 Hz, while c2 was close to 0. This further supports classifying C1 as L*HL instances instead of L*H instances. The contour range of C1 is wide since the rise ranges over more than 4 semitones, and the fall over more than 3 semitones. We therefore interpret C1 as corresponding to L*HL accents with high maturity. C2 corresponds to L*H accents because c2=0, i.e., there is no fall after the peak, and c1>20 Hz. This is typical for accents described as rising accent. As the contour range is more than 4 semitones for the rise, the instances in this cluster can be categorized as L*H accents with wide contour range and therefore

no.	size	b	c1	c2	shape	range	GToBI(S)
C1	31	0.75	30.24	10.16	c1>c2 ⇒ rise	narrow	L*H
C2	14	0.06	4.90	49.62	c1<c2 ⇒ fall	narrow	H*L
C3	9	0.66	110.60	14.15	c1>c2 ⇒ rise	wide	L*H

Table 2: Clusters at **9 months**

no.	size	b	c1	c2	shape	range	GToBI(S)
C1	4	0.79	118.3	2.04	c1>c2 ⇒ rise	wide	L*H
C2	2	0.17	85.99	68.96	c1,c2>30 ⇒ rise-fall	mid	L*HL
C3	5	0.77	12.57	11.35	c1,c2<20		
C4	2	0.52	22.54	93.07	c1<c2 ⇒ fall	wide	H*L
C5	11	0.33	14.78	14.43	c1,c2<20		

Table 3: Clusters at **12 months**

high maturity. C3, finally, corresponds to H*L with a narrow falling contour shape (less than 3 semitones) and therefore low maturity.

At the age of 9 months, we again observe three different clusters (Table 2). For C1, c1>c2, with the range of the rise less than 4 semitones indicates that this cluster corresponds to L*H accents with low maturity. C2 at the age of 9 months can be characterized as falling (c1<c2) with a narrow contour range and thus corresponds to H*L with low maturity. C2 exhibits a very early peak in the accented syllable (b=0.06). A pitch accent at the beginning of a syllable is not optimal for tonal perception [11]. However, since syllable onsets at this age typically consist of one consonant only, and since the fall extends over almost 50 Hz, we assume that these accents are still perceived as narrow falls corresponding to H*L accents with low maturity. C3 can be interpreted as L*H with high maturity: c1 is greater than c2 by more than 90 Hz indicating a rise with wide range.

At the age of 12 months, we can identify five different clusters (Table 3). We assume that two clusters (C3, C5) do not represent pitch accents since both c1 and c2 are less than 20 Hz. Applying the same criteria as before, the remaining clusters can be interpreted as L*H with high maturity (C1), L*HL with mid maturity (C2), and H*L with high maturity (C4). However, the peak alignment for C2 is unusually early in the syllable compared to the adult data [8].

At the age of 14 months, seven different clusters can be found (Table 4). Four of these clusters are characterized as falls (C1, C3, C4, C6), two as rise-falls (C5, C7), and one as unac-

no.	size	b	c1	c2	shape	range	GToBI(S)
C1	12	0.44	5.70	62.01	c1<c2 ⇒ fall	narrow	H*L
C2	10	0.57	17.5	17.56	c1,c2<20		
C3	13	0.16	0.72	35.25	c1<c2 ⇒fall	narrow	H*L
C4	9	0.03	5.74	95.62	c1<c2 ⇒ fall	narrow	H*L
C5	6	0.42	77.88	66.28	c1,c2>30 ⇒ rise-fall	mid	L*HL
C6	3	0.03	24.37	181.39	c1<c2 ⇒ fall	wide	H*L
C7	2	0.47	171.84	108.80	c1,c2>30 ⇒ rise-fall	wide	L*HL

Table 4: Clusters at **14 months**

no.	size	b	c1	c2	shape	range	GToBI(S)
C1	66	-0.06	3.07	106.16	c1<c2 ⇒fall	wide	H*L
C2	22	0.50	82.58	10.18	c1>c2 ⇒rise	wide	L*H
C3	38	0.28	70.47	95.73	c1,c2>30 ⇒rise-fall	wide	L*HL
C4	45	0.41	7.43	126.58	c1<c2 ⇒fall	wide	H*L
C5	57	0.64	12.22	26.77	c1<c2 ⇒fall	narrow	H*L
C6	109	0.15	4.45	46.32	c1<c2 ⇒fall	narrow	H*L

Table 5: Clusters at **18 months**

no.	size	b	c1	c2	shape	range	GToBI(S)
C1	13	-0.11	0	42.77	c1<c2 ⇒fall	narrow	H*L
C2	24	0.32	7.29	48.29	c1<c2 ⇒fall	narrow	H*L
C3	7	0.68	21.32	19.00	c1≈c2		H*
C4	11	0.37	6,21	99.95	c1<c2 ⇒fall	wide	H*L
C5	7	0.86	102.81	1.07	c1>c2 ⇒rise	wide	L*H
C6	7	0.23	82.90	0	c1>c2 ⇒rise	wide	L*H

Table 6: Clusters at **20 months**

cented syllables (C2) since c1 and c2 are less than 20 Hz. The four clusters built for fall mainly differ in their alignment of the peak. C1 has the peak in the middle of the syllable, while C3 exhibits an earlier peak. As the contour range is narrow in both cases, we classify these contours as H*L with low maturity. The peaks of C4 and C6 are both at the beginning of the accented syllable, but the fall in C6 has a higher amplitude. But as stated before tonal movements through areas of maximum new spectral information and intensity change as in syllable onsets might be perceived as level tone [11], particularly in the case of C4, where the fall extends over less than 2 semitones. In the case of C6, we are certain that even with the suboptimal alignment, the contours are perceived as falls since for this cluster, the range of the fall is more than 10 semitones. Therefore, we only tentatively classify C4 as H*L with low maturity, while C6 corresponds to H*L with high maturity. C5 is characterized by a wide range for the rise but only a narrow range for the fall, indicating L*HL with mid maturity. C7 finally has a wide range for both rise and fall, indicating L*HL with high maturity.

At the age of 18 months six different clusters are identified (Table 5). Four clusters are correspond to falling contours (C1, C4, C5, C6). For C1, the alignment of the peak is not optimal for perception, but again, the range extends over more than 5 semitones. We therefore classify these accents as mature H*L. C4 is interpreted as H*L with high maturity, while C5 corresponds to H*L with low maturity. C6 can only tentatively be interpreted as H*L with low maturity since the peak is rather early in the syllable and the range is narrow. C2 is characterized by rising contours with a wide range, thus indicating mature L*H, while C3 is interpreted as mature L*HL.

At the age of 20 months six clusters are found (Table 6). For C3, c1 and c2 have similar values of approximately 20 Hz, thus based on the criterion introduced above, we cannot assign a clear contour shape to these accents. However, these values are too high to dismiss these syllables as unaccented. In terms of GToBI(S), these accents could be classified as H* accents, which for adult data typically exhibit lower c1 and c2 values

no.	size	b	c1	c2	direction	range	GToBI(S)
C1	23	0.50	86.44	45.08	c1,c2>30 ⇒rise-fall	mid	L*HL
C2	50	0.35	7.29	59.40	c1<c2 ⇒fall	narow	H*L
C3	30	0.02	7.29	112.65	c1<c2 ⇒fall	wide	H*L

Table 7: Clusters at **22 months**

than the bitonal accents [8]. Since H* accents in adult data are not characterized by a clear rising or falling movement, we cannot evaluate their contour range. Thus, it is also not possible to compare the child contour to adult-like contours in this case. The remaining clusters correspond to falls (C1, C2, C4) or rises (C5, C6). We interpret these as H*L with low maturity (C1, C2), H*L with high maturity (C4), and L*H with high maturity (C5, C6). Again, it should be noted that the tonal perception of the contours in C1 might be compromised because of the very early peak.

At the age of 22 months we can identify three clusters (Table 7) corresponding to mid mature L*HL (C1), H*L with low maturity (C2), and H*L with high maturity (C3).

4. Discussion

In contrast to child speech data for American English [1], our data suggest that rising contours and rise-fall contours are the first contours that are consistently used by children by the age of 7 months already. At 7 months, we do observe falling contours (Table 1, C3), but these are realized with a narrow range, while the contours in the two clusters corresponding to rise and rise-fall (Table 1, C2 and C1) are realized in an adult-like manner with a wide range. It is only at 12 months that we find a first falling cluster corresponding to mature H*L (Table 3, C4). After that, the proportion of clusters interpreted as H*L increases, but the number of different clusters attributed to H*L accents shows that these are still produced with high variability. We have discussed above that this variability is due to different peak alignment or amplitude range in the different clusters. Based on the clustering results, we conclude that the variability starts to decrease at the age of 20 months, arriving at only two clusters for H*L at the age of 22 months. Preliminary data from later recordings of the same child, which we have not discussed here, indicate that this tendency persists even to the age of 33 months.

It seems that while the frequency of falling contours increases, that of purely rising contours decreases. While at 9 months, 75% of the accents were interpreted as rising accents, at 14 months, none of the clusters can be interpreted as a rise cluster (although rise-falls do occur as two clusters can be interpreted as L*HL). At 18 months, only one cluster is a rise cluster, corresponding to 6% of the accents, and at 20 months, two rise clusters can be found (20% of the accents). At 22 months, finally, no cluster can be interpreted as purely rising. However, cluster C1 in Table 7, which corresponds to L*HL with mid maturity, might include some accents that could also be interpreted as rises. We classified cluster C1 as mid mature because the fall amplitude was smaller than usually observed for L*HL in adult speech. In some cases the amplitudes may have been so low that these accents could also be classified as L*H.

The results presented in this study are results from a case study on German speech acquisition involving only one child. Future work on more data from the Stuttgart Child Language Corpus [10] will examine whether these results can be confirmed using data from other children.

5. Conclusion

This case study was intended to verify that our method, viz. the parametrization of F0 contours in combination with a clustering technique, is suitable for examining both babbling and meaningful child speech. We have shown that from the cluster results, we can derive contour shapes and contour range, which permits comparing our results to studies using traditional contour-based descriptions. On the other hand, the PaIntE parameters allow for assigning realizations of accents to adult ToBI categories.

The advantage of our method is that using the PaIntE parametrization, we can capture fine phonetic detail such as peak alignment and rise and fall amplitudes in children's realizations of accent contours. Using clustering methods to further analyze these data, we can assess the variability of the child's production of intonation contours. This method can be applied to both babbling and more complex multi-word utterances, which is favorable for longitudinal studies of intonation in child speech because we can apply the same method over the course of the study even as children go through different developmental stages from pre-linguistic utterances to multi-word utterances.

6. Acknowledgements

This study is part of the project *Tonal and temporal development of prosodic categories of German speaking children* funded by the German Research Foundation (DFG, Grant MO 597/3-1).

7. References

[1] H. L. Balog and D. Snow, "The adaption and application of relational and independent analyses for intonation production in young children," *Journal of Phonetics*, vol. 35, pp. 118–133, 2007.

[2] D. K. Oller, S. N. Iyer, E. H. Buder, K. Kwon, L. Chorna, and K. Conway, "Diversity and contrastivity in prosodic and syllabic development," in *Proceedings of ICPhS (Saarbrücken)*, J. Trouvain and W. J. Berry, Eds., 2007, pp. 303–308.

[3] J. Hirschberg and M. E. Beckman, *The ToBI annotation conventions*, 1994.

[4] P. Prieto and M. d. M. Vanrell, "Early intonational development in Catalan," in *Proceedings of ICPhS (Saarbrücken)*, J. Trouvain and W. J. Barry, Eds., 2007, pp. 309–314.

[5] A. Chen and P. Fikkert, "Intonation of early two–word utterances in Dutch," in *Proceedings of ICPhS (Saarbrücken)*, J. Trouvain and W. J. Barry, Eds., 2007, pp. 315–320.

[6] P. A. Halle, B. de Boysson-Bardies, and M. M. Vihman, "Beginnings of Prosodic Organization: Intonation and Duration patterns of Disyllables produced by Japanese and French Infants," *Language and Speech*, vol. 34, no. 4, pp. 299–318, 1991.

[7] G. Möhler and A. Conkie, "Parametric modeling of intonation using vector quantization," in *Proc. 3rd ESCA Workshop on Speech Synthesis*, 1998, pp. 311–316.

[8] A. Schweitzer, "The tonal dimension of perceptual space," Ph.D. dissertation, University of Stuttgart, in preparation.

[9] J. Mayer, "Transcription of German intonation – the Stuttgart system," University of Stuttgart, Tech. Rep., 1995.

[10] B. Lintfert, "Phonetic and phonological development of stress in German," Ph.D. dissertation, University of Stuttgart, 2009.

[11] D. House, "Differential perception of tonal contours through the syllable," in *Proceedings of ICSLP*, vol. 1, 1996, pp. 2048–2051.