Describing the development of intonational categories using a target-oriented parametric approach

Britta Lintfert[1,2], *Bernd Möbius*[1]

[1]Department of Computational Linguistics and Phonetics, Saarland University, Germany
[2] Institute of Natural Language Processing, University of Stuttgart, Germany

britta.lintfert@ims.uni-stuttgart.de, moebius@coli.uni-saarland.de

Abstract

In this paper we analyze the relation between adults' intonational categories as described in the ToBI framework and children's intonation contours, using a parametric approach and cluster evaluation methods. In the field of prosody, an increasing number of studies on the development of intonation apply the intonational categories of adult speech described as a sequence of high (H) and low (L) tones to child speech [1, 2]. However, the categories described by ToBI or by its language-specific variants are developed for adult speakers. Instead of imposing adult category representations on the description of developmental stages in L1 intonation acquisition, we propose and validate a parametric approach and cluster evaluation methods [3, 4] that are applicable to both adult and child speech. First, we show that clusters of parametrized contours obtained from German adult-directed and child-directed speech correlate well with German ToBI categories. We then assess how well clusters at different stages of intonation acquisition correspond to adult target categories. Our results indicate that the proposed methodology is capable of demonstrating, in qualitative and quantitative terms, the continuous development of intonational categories at early ages towards adult target categories.

Index Terms: development of intonation, F0 parametrization, GToBI(S), clustering, adult targets

1. Introduction

To describe the development of intonation one common approach is to compare the child's productions to a mature model (i.e., the adult model), mainly within the ToBI framework [5, 6, 7]. ToBI approaches analyze intonation contours as sequences of (possibly categorical) intonation events, where each event can be decomposed into high and low pitch targets which are aligned with the syllable structure. However, the categories posited 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 speechlike productions are already capable of consistently using the categories posited by intonational theory.

Against this background, to find categories of intonation even in pre-linguistic production of child speech we have suggested an automatic method for describing the shape of the F0 contour [3]. We proposed to parametrize F0 contours in the vicinity of accented syllables by PaIntE approximation [8]. Groups of similar contours can then be identified by *K*-means clustering, reasoning that different clusters may be interpreted as different intonational categories. Results on adult data of child-directed speech showed a much better than chance correspondence between adult clusters and GToBI(S) categories [4].

In this paper, we validate the idea of mapping clusters to ToBI categories on child speech at different developmental stages. We also compared the child clusters to adult target ToBI categories to show a developmental pattern of intonation contours. We therefore extend the methodology described in [4] to speech produced by children aged between 1 and 8 years. We compare the intonation contours produced at different ages to the adult target form. This method facilitates the description of a developmental pattern, evolving towards adult targets, as indicated by accuracies increasing with age.

2. Method

2.1. Corpus and data preparation

For this study we used longitudinal data from 9 children. The recordings are part of the Stuttgart Child Language Corpus [9] and took place at the children's homes in familiar play situations with their mothers while looking at picture books or playing with toys. Thus the data represent spontaneous speech productions. German child-directed (CDS) as well as adult-directed speech (ADS) of the mothers of the children were also recorded and analyzed (see Table 1).

The recordings were made with two wireless microphones AKG CK 97-L and a Marantz PMD670 Flash Recorder with a 2 GB CF-Card at a sampling rate of 48 kHz. All recordings were transferred to a computer workstation, downsampled to 16 kHz and manually annotated on the segment, syllable and word level. The children's utterances were manually annotated with respect to perceived prominence. Syllables classified as prominent were then additionally coded according to GToBI(S) pitch accent categories. Note that this coding only served as a reference for comparing the children's production of F0 contours with adult targets. It does not imply an interpretation of child speech in terms of adult categories. The parametrization was performed for the prominent, and thus potentially accented, syllables only. GToBI(S) is an adaptation of ToBI to German and 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.

Inter-observer reliability was assessed on 10% of the annotated data. Inter-observer agreement on the segmental and syllable levels was 94.5%, 88.3% on the word level, and 77.8% on the prosodic level.

2.2. PaIntE parametrization

PaIntE stands for "Parametrized Intonation Events" [8] and was originally developed for F0 modeling in speech synthesis.

age	mean age (in months)	number of accent tokens	number of children
1;0	13.36	354	3
2;0	20.04	1087	4
3;0	35.23	1848	5
4;0	47.55	2490	3
5;0	59.05	147	2
6;0	71.26	690	2
7;0	83.10	290	1
8;0	93.71	539	2
ADS		191	
CDS		4512	

Table 1: Overview of analyzed accents

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.

In contrast to other F0 parametrization or stylization approaches, PaIntE attempts to directly model properties of F0 contours that have been claimed to be linguistically meaningful. For instance, parameters *c1* and *c2* are intended to capture the amplitude of the pitch movement. Parameter *b* quantifies the alignment of the peak with the syllable structure. Pitch movement excursion and peak alignment are tonal correlates of prominence and pitch accent type, respectively.

2.3. Cluster analysis

K-means clustering is a hard clustering method which partitions the data into *k* clusters. The number of clusters *k* has to be specified beforehand. Each cluster is defined by its centroid: each observation belongs to the cluster with the nearest centroid.

For the experiments presented here, we used R's [11] `kmeans` function, which by default implements the Hartigan-Wong method [12]. We ran `kmeans` with 100 random starts, varying the number of clusters from 2 to 9, to cluster the data. We used all six PaIntE parameters as attributes, which were z-scored to eliminate speaker-specific and age-specific effects of pitch range and key and to match them with respect to scaling, which ensures that all parameters have approximately equal importance in clustering. The labeled accents were not used as attributes for clustering. We used up to 9 clusters as we assume that each pitch-accent category can have more than one cluster depending on the alignement of the peak in the syllable [3].

For all experiments we describe the correspondence between the clusters and the manually annotated GToBI(S) [10] category by calculating the classification accuracy. Since the clustering identifies groups of similar contours, it stands to reason that the clusters might represent typical instances of specific GToBI(S) events [13].To this end, each cluster was assigned the GToBI(S) accent which occurred most frequently in the cluster. We then evaluated for how many accents their manually annotated "true" category did indeed correspond to the category which had been assigned to their cluster [4]. The score obtained in this evaluation can be interpreted as classification accuracy: clusters are used to classify GToBI(S) accents based on the observed PaIntE parameters, with the score indicating the percentage of correct decisions and correspondence correlating with accuracy. The accuracy that can be reached if the accents are classified as belonging to the most frequent GToBI(S) category is indicated for each experiment as baseline accuracy.

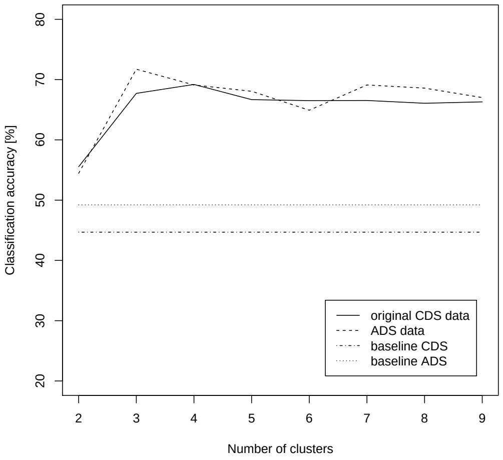

Figure 1: Classification accuracies for 2 to 9 clusters obtained from child-directed (CDS) clustering data (solid line) and adult-directed data (ADS) (dashed line). The baselines are the relative frequencies of the most frequent accents in CDS (dot-dashed line) and ADS (dotted line) data.

3. Results

3.1. Mapping adult clusters to GToBI(S) categories

To analyze whether the mothers vary in the production of categories depending on the speech task, we compared the cluster results obtained from child-directed and adult-directed data, respectively. To assess whether the cluster-to-category assignment is indeed similar for both data, we applied the clustering from child-directed speech data to adult-directed speech data by assigning each datapoint of the ADS data to the nearest cluster center in the CDS data. We then evaluated, in how many cases the category assigned to the cluster based on CDS data matched the manually annotated "true" GToBI(S) category.

Both classification accuracies, the one obtained directly from child-directed clustering data (CDS) and the one from adult-directed data (ADS) using clusters and cluster-to-category assignment from the child-directed data, are depicted in Figure 1, as a function of number of clusters. The solid line indicates the classification accuracy for child-directed original clustering data. As can be seen, accuracies of between approx. 65 and 70% are reached. The accuracies for adult-directed data (dashed line) are generally higher, except when using six clusters. Based on these results it can be concluded that the clusters obtained for adult-directed speech are indeed similar to those in child-directed speech. For comparison, the baseline accuracies, i.e. the accuracy that can be reached if one simply classifies all accents as belonging to the most frequent GToBI(S) category, are indicated by the dot-dashed (CDS) and dotted (ADS) lines. They are at 44.7% and 49.2%, respectively. The results thus indicate a much better than chance correspondence between clusters and GToBI(S) categories, and the categories produced within adult-directed speech correspond well with the

age	classification accuracy		baseline accuracy
	mean	sd	
1;0	64.8	1.9	61.6
2;0	61.0	1.8	58.4
3;0	59.5	3.6	50.8
4;0	64.9	1.7	54.6
5;0	66.9	3.8	42.6
6;0	65.6	1.8	46.0
7;0	59.2	4.9	42.6
8;0	64.6	1.1	40.5

Table 2: Classification and baseline accuracies for the correspondence between the clusters and GToBI(S) categories.

categories produced in child-directed speech. There is no difference in intonational categories as a function of the speech task.

3.2. Development of intonational categories

To describe the development of categories we investigated the correspondence between the clusters of the child data at each age and the baseline accuracy. We therefore applied the clustering method to prosodically annotated data for children aged between about 1 and 8 years. For a better overview of possible developmental patterns we grouped the data depending on age and describe the development of categories for the children (see Table 1). We investigated the correspondence between the clusters and GToBI(S) [10] categories. To this end, each cluster was assigned the GToBI(S) accent which occurred most frequently in each cluster. We then evaluated for how many accents their manually annotated "true" category did indeed correspond to the category which had been assigned to their cluster. The classification accuracies (mean and standard deviation) as well as the baseline accuracies for all groups are given in Table 2, for the children between 4 and 8 years both are also depicted in Figure 2 as a function of number of clusters. Note that the baselines for 5 and 7 years are identical (42.6).

At about one year, there is almost no difference between classification accuracy (64.8) and baseline (61.6). The results thus indicate not a better than chance correspondence between clusters and GToBI(S) categories. Nearly the same holds for the children at about two years with a classification accuracy of 61.0 and a similar high baseline accuracy of 58.4. The picture change at about three years of age. The difference between classification and baseline increases and, more importantly, the baseline decreases but remains at chance level until the age of about four. The baseline falls below chance level at about five years of age. Based on these results we conclude that from about 5 years of age onward the children produced intonation categories that correlate with GToBI(S) categories. The data obtained for children older than four years indeed showed a much better than chance correspondence between clusters and GToBI(S) categories.

3.3. Mapping adult targets and child categories

If the clustering serves to identify "categories" in intonation contours, then we would expect that for the children data introduced above, these categories correspond well to the adult GToBI(S) categories. To assess whether the cluster-to-category assignment is indeed similar for the children data comparable to the adult targets, we applied the clustering obtained on child-directed speech data to the children data at each age by assigning each datapoint of the children data to the nearest cluster center obtained on CDS data. We then evaluated, analogously to the evaluation of CDS data, in how many cases the category assigned to the cluster based on CDS data matched the manually annotated "true" GToBI(S) category.

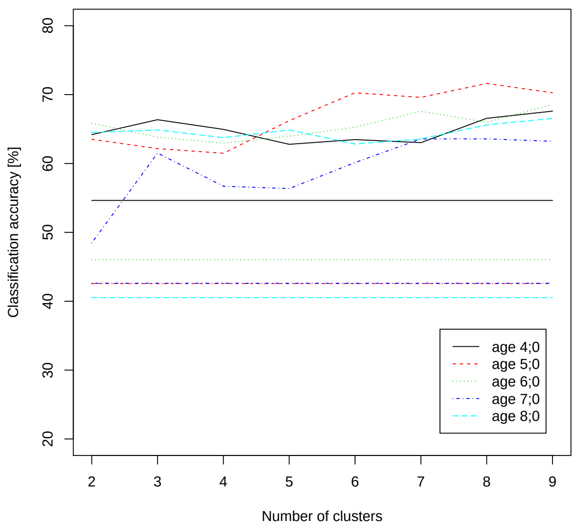

Figure 2: Classification accuracies for 2 to 9 clusters for clustering data of children between 4 and 8 years of age. The baselines are the relative frequencies of the most frequent accents in the child data.

Until the age of about four years the classification accuracy based on the mapped clusters were even worse than the baseline accuracy, i.e., the accuracy that can be reached if one simply classifies all accents as belonging to the most frequent GToBI(S) category (Table 3). The categories produced by the children until the age of 4 years do not correspond well with the categories produced in child-directed speech. These results indicate that with the beginning of speech-like productions children are not yet capable of consistently using the categories as posited by intonation theory. At the age of about 4 years this picture changes, as indicated by the increase of classification accuracy based on the mapped clusters and the decrease of baseline accuracy (Figure 3). Moreover, the accuracies reached for mapped clusters are worse than those without mapping of adult target categories. A tendency toward more adult-like intonational categories can be observed. Based on these results we conclude that from about 5 years of age the children produced categories comparable to those in adult intonation.

4. Conclusion

This study aimed to verify that our proposed methodology, viz. the parametrization of F0 contours in combination with a clustering technique, is suitable for identifying intonational "categories". The results of Experiment 2 showed that until the age of about 5 years, there is almost no difference between the classification accuracy and the baseline. Our interpretation is that that the categories produced by younger children cannot be described adequately by means of GToBI(S) categories. With increasing age the accuracies increase, too, and at about 5 years

age	classification accuracy		baseline accuracy
	mean	sd	
1;0	55.9	3.4	61.8
2;0	51.6	2.7	58.4
3;0	67.4	2.9	63.4
4;0	59.7	2.7	54.7
5;0	64.5	3.5	42.9
6;0	62.6	5.3	64.1
7;0	60.2	4.3	42.8
8;0	61.2	3.9	40.5

Table 3: Classification and baseline accuracies for mapping adult targets and children categories.

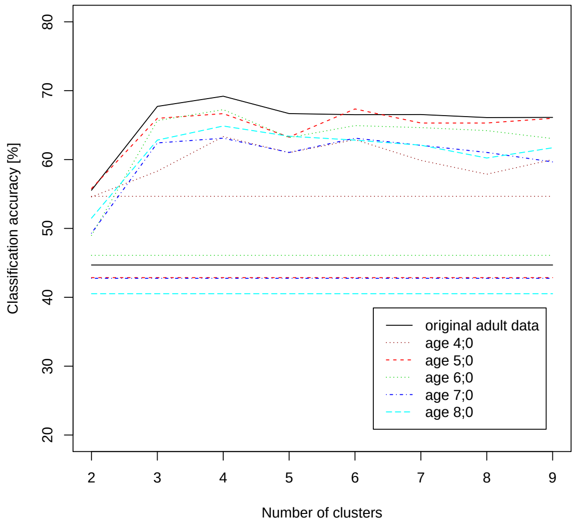

Figure 3: Classification accuracies and baselines for 2 to 9 clusters, based on child-directed clustering data (solid line) and child data between 4 and 8 years of age. Note that the baselines for 5 and 7 years are similar.

of age children tend to produce categories similar to those assumed by GToBI(S), resulting in a lower baseline and higher classification accuracies. In Experiment 3 the categories produced by the children were mapped onto adult targets to identify possible GToBI(S) categories. As expected from the results of experiment 2, until the age of five the produced categories do not correspond well to the underlying adult targets. Children older than 5 years are able to produce categories similar to those of adult speakers, corresponding well with GToBI(S) categories. For younger children, however, GToBI(S) categories cannot describe the produced intonation categories adequately.

The advantage of the proposed method is that by using the PaIntE parametrization, it can capture fine phonetic detail such as peak alignment and rise and fall amplitudes in children's realizations of intonation contours. Using clustering methods to further analyze the data, we can assess the variability of the children's production of intonation contours and classify the maturity of the contours depending on the children's age. We can also capture the variability in peak alignment as well as amplitude of rises and falls in the children's production of intonation contours and can describe which categories children produce and how they differ from adult categories. This method can be applied to all stages of L1 intonation acquisition but also to adult speech. This is favorable for longitudinal studies of intonation in child speech, as 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. An extension of the method by including temporal parameters is envisaged in future work.

The proposed method is an automatic approach and can give comparable results independent of the theoretic analysis framework. The clustering method can also include additional variables, such as those related to discourse structure.

5. Acknowledgements

Special thanks to Antje Schweitzer who wrote the analysis scripts. 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).

6. References

[1] A. Chen, "The developmental path to phonological focus-marking in dutch." in *Prosodic categories: Production, perception and comprehension*, . P. P. S. Frota, E. Gorka, Ed. Dordrecht: Springer., 2011, pp. 93–109.

[2] P. Prieto, A. Estrella, J. Thornson, and M. Vanrell del Mar, "Is prosodic development correlated with grammatical and lexical development: Evidence from emerging intonation in catalan and spanish," *Journal of Child Language*, vol. 21, pp. 1–37, 2011.

[3] B. Lintfert, A. Schweitzer, L. Wolski, and B. Möbius, "Quantifying developmental changes of prosodic categories," in *Proceedings of Speech Prosody 2010*, Chicago, Illinois, 2010.

[4] B. Lintfert, A. Schweitzer, and B. Möbius, "A parametric approach to intonation acquisition research: Validation on child-directed speech data," in *Proceedings of Interspeech 2011, Florenz.*, 2011.

[5] J. B. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, MIT, 1980.

[6] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labelling English prosody," in *Proceedings of the International Conference on Spoken Language Processing (Banff, Alberta)*, vol. 2, 1992, pp. 867–870.

[7] J. Pitrelli, M. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP, Yokohama)*, 1994, pp. 123–126.

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

[9] B. Lintfert, *Phonetic and phonological development of stress in German*. Doctoral dissertation, Universität Stuttgart, 2009. [Online]. Available: http://elib.uni-stuttgart.de/opus/volltexte/2010/5424/

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

[11] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[12] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100–108, 1979.

[13] A. Schweitzer, *Production and Perception of Prosodic Events—Evidence from Corpus-based Experiments*. Doctoral dissertation, Universität Stuttgart, 2011. [Online]. Available: http://elib.uni-stuttgart.de/opus/volltexte/2011/6031/