A parametric approach to intonation acquisition research: Validation on child-directed speech data

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

[1]Institute of Natural Language Processing, University of Stuttgart, Germany
[2] Department of Computational Linguistics and Phonetics, Saarland University, Germany
{britta.lintfert,antje.schweitzer}@ims.uni-stuttgart.de, moebius@coli.uni-saarland.de

Abstract

This paper validates a parametric approach to intonation acquisition research [1] using child-directed speech data. An advantage of this approach is that it can be used for studying child speech as well as adult speech. Within the field of prosody acquisition it reconciles independent approaches to child prosody with ToBI-based approaches. In this paper we substantiate this claim by showing that clusters of parameterized contours obtained from German child-directed speech correlate with GToBI(S) categories, and by elaborating how, alternatively, the parameters can be mapped to properties that are relevant in independent approaches.

Index Terms: intonation, F0 parameterization, clustering, child-directed speech

1. Introduction

In the field of prosody acquisition two different approaches are common to describe the development of intonation [2]: the *independent* and the *relational* approach. In an independent analysis of intonation [3, 4] the child's productions are not compared to mature models. Intonation contours are described with reference to properties such as direction (i.e., falling or rising), accent range (i.e., amplitude of pitch change), and complexity (e.g., changes in direction measured in semitones). For instance [5] report a description of the developing patterns based on these measurements. In contrast, in a relational analysis, the child's productions are compared to a mature model (i.e., the adult model). A common model for describing mature intonation is the ToBI framework [6, 7, 8]. 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. Beyond the identification of pitch targets and their coarse alignment with the syllable structure, finer aspects of the phonetic realization of these events, such as amplitude of the pitch movements, or exact peak alignment within syllables, are not analyzed in the ToBI framework. 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 meaningful speech are already capable of consistently using the categories posited by intonational theory.

Against this background we have suggested [1] an automatic method for analyzing F0 contours which is compatible with both approaches. We proposed to parameterize F0 contours in the vicinity of accented syllables by PaIntE approximation [9] (see section 2.2). This yields several parameters which describe the shape of the F0 contour around the accented syllable. We then identified groups of similar contours by *K*-means clustering, reasoning that different clusters may be interpreted as different intonational categories. First results on data from one child showed that more clusters could be characterized as rises or rise-falls than as falling accents until the age of 1;1 [1]. After that the proportion of clusters interpreted as falling accents increased but the number of different clusters attributed to falling accents showed that these were still produced with high variability mainly due to different peak alignment.

This methodology can be used for studying the development of intonation in both babbling and meaningful child speech. Furthermore, our method is compatible with the independent approach: properties such as accent range, direction, or complexity, can be derived from the PaIntE parameters. For comparison of the clusters with the ToBI approach, the clusters can be mapped to ToBI categories. In this paper, we validate the idea of mapping clusters to ToBI categories on German child-directed speech data. The quality of the mapping from clusters to GToBI(S) [10] categories is evaluated in terms of classification accuracy to assess comparability objectively.

2. Method

2.1. Participants and data collection

For this study we examined German child-directed speech of two female adult speakers, AD and AL. The data amounts to 2635 accented syllables (AD 1613, AL 1022). The recordings are part of the Stuttgart Child Language Corpus [11] and took place at the women's homes in familiar play situations with their children aged between 3;4 and 4;6 while looking at picture books or playing with toys. Thus the data represent spontaneous child-directed productions. The recordings were made with a wireless microphone 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, and manually prosodically labeled according to GToBI(S) [10].

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. For L*H and H*L, allotonic variants exist, for instance, monotonal L* for L*H, or monotonal H* for H*L. The main differences to GToBI[12] are (i) that with regard to their function in discourse, H* and L* are claimed to be equivalent to H*L and L*H, respectively; and (ii) that GToBI(S) does not distinguish between L+H* and L*+H accents—the former would correspond to just an H* or to a sequence of an allo-

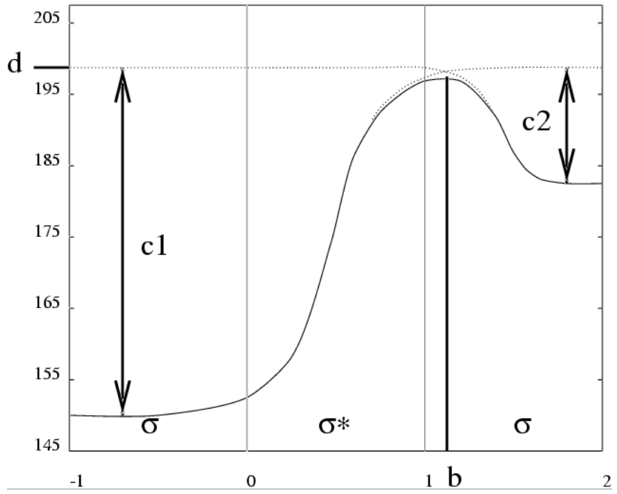

Figure 1: Schematic of the PaIntE approximation function, reproduced from [9]. The approximation window represents three syllables, where the accented syllable is indicated by the asterisk (σ*). Peak height is determined by parameter *d*, amplitudes of rise and fall correspond to parameters *c1* and *c2*, respectively, and peak alignment depends on the *b* parameter.

tonic variant of a preceding H*L followed by an H* accent, the latter would correspond to an L*H accent in GToBI(S).

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 parameterization

PaIntE stands for "Parameterized Intonation Events" [9] 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 parameterization 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 sigmoids. A schematic of the function is given in Figure 1. The time axis is normalized to the lengths of the syllables, e.g., the peak is at the beginning of the accented syllable if *b*=0, and at its end if *b*=1.

In contrast to other F0 parameterization 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.

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 [13] `kmeans` function, which by default implements the Hartigan-Wong method [14]. We used `kmeans` to cluster AD's data, varying *k* from 2 to 9, with 30 random starts. In this setting `kmeans` clusters the data 30 times for each *k*, using different initial cluster centers, and picks the clustering for which the sum of squares from points to the assigned clusters is minimal. We used all six PaIntE parameters as attributes; however, we converted parameters *c1* and *c2*, which specify the amplitude of the F0 movement, from Hertz to semitones in order to model human perception more closely [15], and to achieve a more direct correspondence to the independent approach, which defines maturity of accent range in terms of semitones [5]. All parameters were then z-scored to eliminate speaker-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.

3. Results

3.1. Comparing PaIntE to independent approaches

For an independent analysis comparable to [5] we can derive properties such as range, direction, and complexity from the cluster centers. High maturity according to [5] in falling accents is indicated if the accent range is greater than 4 semitones, in rising accents if the accent range is greater than 3 semitones. These properties can be directly derived from parameters *c1* and *c2*, using *d* as a reference. Second, accent direction can be compared to that in mature productions by comparing parameters *c1* and *c2*, i.e., by the relation between rise and fall amplitudes. An overall fall is given for *c2*>*c1* and an overall rise for *c1*>*c2*. Parameters *c1* and *c2* can also capture complexity to some extent. A simple heuristic could be that if both *c1* and *c2* exceed a certain threshold, say, 1 semitone, the contour can be characterized as rise-fall, indicating greater complexity than rise-only or fall-only contours.

Please note that for an independent analysis, the cluster analysis is not immediately necessary. Clustering groups similar realizations together. In each such group of similar realizations, the centroid can be interpreted as the "prototypical" realization. Thus, range, accent direction or complexity can either be regarded separately for each instance or they can be regarded for the centroids only, with each centroid representing a cluster of similar realizations.

3.2. Comparing PaIntE to ToBI approaches

[16] has shown on a prosodically annotated corpus of a male German speaker that the PaIntE parameter distributions of the GToBI(S) accents indeed capture the defining properties of German pitch accents. For instance, H*L accents usually have their peak in the middle of the accented syllable, which is consistent with a high pitch target associated with this syllable, as claimed by GToBI(S). In contrast, in non-final L*H accents, the peak can occur in the middle of the post-accented syllable, again in line with GToBI(S) expectations. Also, falling (H*L) accents usually have greater *c2* than *c1* values (i.e., the amplitude of the falling sigmoid is greater than that of the rising sigmoid); vice versa for rising (L*H) accents.

These observations motivate the idea that there is a correspondence between GToBI(S) events and PaIntE parameters. Since the clustering identifies groups of similar contours, it stands to reason that the clusters might represent typical instances of specific GToBI(S) events.[1]

Thus, there are two ways to investigate the relationship be-

[1] This idea was investigated in detail by [16], however, with slightly different aims and different attributes.

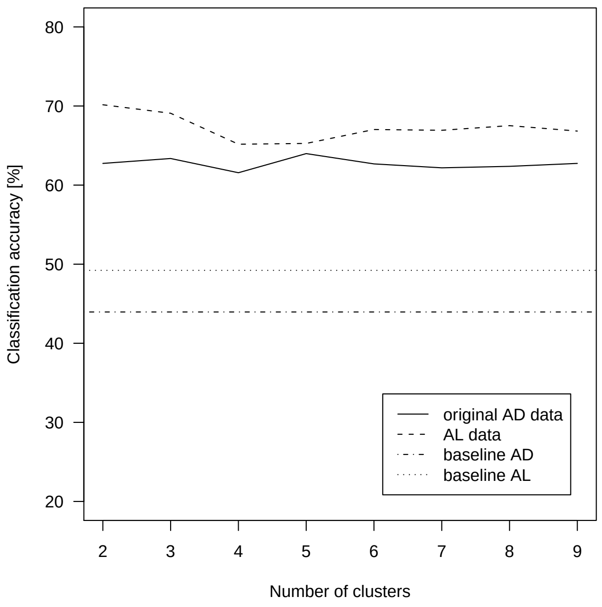

Figure 2: Classification accuracies for 2 to 9 clusters on the clustering data (solid line) and on AL's data (dashed line). The baselines are the relative frequencies of the most frequent accents in AD's (dot-dashed line) and AL's (dotted line) data.

tween PaIntE parameters and GToBI(S) events. One is to detect a direct mapping from the parameters to the GToBI(S) events. The other way is to investigate the mapping from clusters to GToBI(S) events, i.e. the cluster analysis is an intermediate step in mapping from parameters to GToBI(S) categories.

We claim that the clustering approach is the more interesting one for analyzing prosody acquisition because it allows to find prototypical realizations in the child's productions independent of the established adult GToBI(S) categories, i.e., it allows to identify "categories" at each developmental stage, and to compare these "categories" to mature categories in a second step. Alternatively, without the intermediate clustering, one would directly compare each production of the child to mature GToBI(S) categories.

We will pursue both approaches here; the former because we consider it the more appropriate way in child prosody; the latter in order to evaluate if the intermediate clustering step obscures the correlation between parameters and categories.

3.2.1. Mapping clusters to categories

If the clustering serves to identify "categories" in intonation contours, then we would expect that for the adult data introduced above, these categories correspond well to the adult GToBI(S) categories. To assess the correspondence between clusters and GToBI(S) categories, we assigned categories to the clusters obtained on AD's data, in line with our claim that the clusters represent different categories. 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. This procedure was also employed in [16]. The score obtained in this evaluation can be interpreted as classification accuracy: if the clusters were used to classify GToBI(S) accents based on the observed PaIntE parameters, this score indicates the percentage of correct decisions.

Please note that this procedure is not strictly unsupervised, as the decision of which category corresponds to a cluster is based on the clustering data and requires the categories to be known beforehand. Of course, we hope that the cluster-to-category assignment captures general properties of the GToBI(S) categories and not only properties specific to the clustering data, so it should hold for other data as well. To assess whether the cluster-to-category assignment is indeed valid for other data, we applied the clustering obtained on AD's data to AL's data by assigning each datapoint of the AL data to the nearest cluster center obtained on AD's data. We then evaluated, analogously to the evaluation of AD's data, in how many cases the category assigned to the cluster based on AD's data matched the manually annotated "true" GToBI(S) category.

Both classification accuracies, the one obtained directly on AD's clustering data, and the one obtained on AL's data using AD's clusters and cluster-to-category assignment, are depicted in Figure 2, as a function of number of clusters. The solid line indicates the classification accuracy obtained on AD's original clustering data. As can be seen, accuracies of between approx. 60 and 64% are reached. Interestingly, the accuracies on AL's data (dashed line) are even higher and can reach 70%. 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 (AD) and dotted (AL) lines. They are at 44.0% and 49.2%, respectively. The results thus indicate a much better than chance correspondence between clusters and GToBI(S) categories.

3.2.2. Mapping parameters to categories

In order to get an impression of the classification accuracy one could obtain by directly mapping from PaIntE parameters to categories, we used WEKA [17] to train classifiers to predict GToBI(S) accents based on the same attributes as we have used for clustering. We tried all 69 learning schemes available in WEKA 3.7.1 which were applicable to the present problem, using the default settings except for IBk (instance-based learning), where the default setting with *k*=1 would have yielded the IB1 scheme. Here, we used *k*=20. The classifiers were trained and evaluated on AD's data using 10-fold cross validation. Figure 3 presents an overview. Classifiers are listed in the order of their performance. We show only every other scheme to save space; the remaining schemes still give an impression of the classification accuracies. The names are the original scheme names in WEKA. The vertical solid line is at 63%, which was approx. the classification accuracy that could be reached in the clustering experiments.

It can be seen that not all learning schemes are equally suitable to the present problem. However, various schemes yield accuracies of greater than 60%, and many of them yield rates that slightly exceed the accuracies obtained in the clustering experiments. Thus the results show that the clustering approach serves to identify groups of similar instances that correspond well to the GToBI(S) categories: classification accuracy is only slightly lower than the accuracy that can be obtained when directly training classifiers to predict the categories. Beyond providing a reference for evaluating the correspondence between clusters and categories, these results indicate the strength of the correspondence between PaIntE parameters and GToBI(s) categories in general, without the intermediate clustering step.

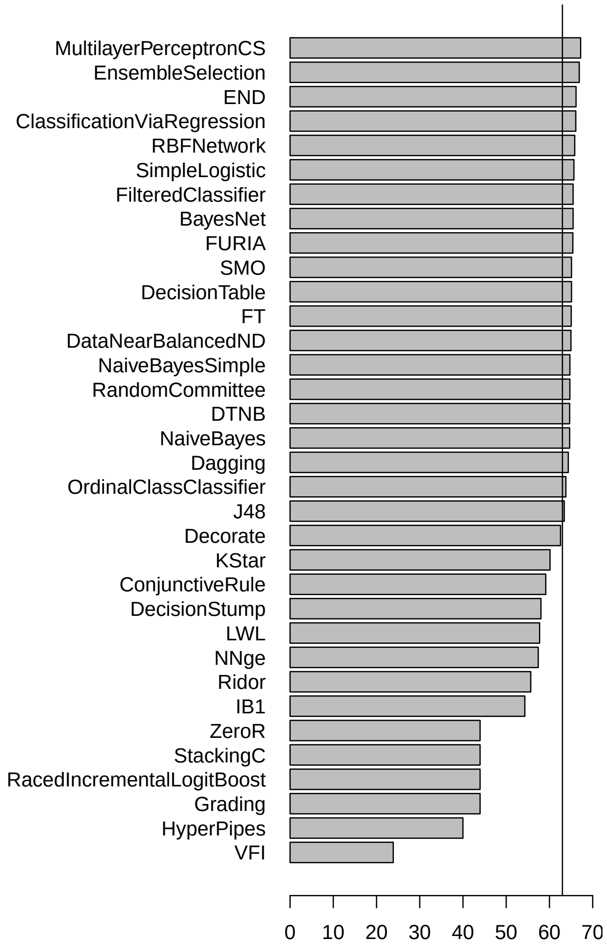

Figure 3: Estimated classification accuracies (in percent) for classifiers trained on AD's data, evaluated by 10-fold cross-validation using WEKA. See text for further details.

4. Conclusion

This study intended to verify that our methodology, viz. the parameterization of F0 contours in combination with a clustering technique, is suitable for identifying intonational "categories". The main motivation for introducing this methodology lies in the fact that it combines advantages of existing frameworks for analyzing child prosody. We can derive range, direction and complexity of the contour either from the cluster results or directly from the PaIntE parameterization, which permits a comparison of our results to findings obtained in the independent approach. Alternatively, we can identify the GToBI(S) accent corresponding to each cluster for compatibility with the relational approach. Using the PaIntE parameterization, we can capture fine phonetic detail such as peak alignment within syllables and rise and fall amplitudes in realizations of accent contours. In our opinion, an interesting advantage of the clustering method is that it attempts to detect "categories" in children's productions, and that the development of these categories can be investigated. These categories can be similar to the adult target form or vary depending on the children's limitations in production. This method can be applied to infant and child speech as well as to adult speech. It is thus suitable for longitudinal studies formally describing the acquisition of intonation in different developmental stages, from pre-linguistic utterances to multi-word utterances. Because the method is also language-independent, it facilitates cross-language intonation studies too.

5. 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).

6. References

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

[2] C. Stoel-Gammon and C. Dunn, *Normal and Disordered Phonology in Children*. Baltimore: University Park Press, 1985.

[3] A. Cruttenden, *Intonation*, 2nd ed. Cambridge, MA: Cambridge University Press, 1997.

[4] D. Crystal, "Prosodic development," in *Language Acquisition*, P. Fletcher and M. Garman, Eds. Cambridge University Press, 1986, pp. 174–198.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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/

[12] M. Grice, S. Baumann, and R. Benzmller, "German Intonation in Autosegmental-Metrical Phonology," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford University Press, 2005.

[13] 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

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

[15] F. Nolan, "Intonational equivalence: An experimental evaluation of pitch scales," in *Proceedings of the International Congress of Phonetic Science*, 2003, pp. 771–774.

[16] 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/

[17] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, USA: Morgan Kaufman, 2005.