Erlangen-CLP: A Large Annotated Corpus of Speech from Children with Cleft Lip and Palate

Tobias Bocklet[1,2]**, Andreas Maier**[2]**, Korbinian Riedhammer**[2]**, Ulrich Eysholdt**[1]**, Elmar Nöth**[2,3]
[1]Department of Phoniatrics and Pediatric Audiology, University Hospital Erlangen, Germany
[2]Pattern Recognition Lab, University of Erlangen-Nuremberg, Germany,
[3]E&C Engineering, King Abdulaziz University, Saudi Arabia
tobias.bocklet@cs.fau.de, andreas.maier@fau.de, korbinian.riedhammer@cs.fau.de,
ulrich.eysholdt@uk-erlangen.de, noeth@cs.fau.de

Abstract

In this paper we describe Erlangen-CLP, a large speech database of children with Cleft Lip and Palate. More than 800 German children with CLP (most of them between 4 and 18 years old) and 380 age matched control speakers spoke the semi-standardized PLAKSS test that consists of words with all German phonemes in different positions. So far 250 CLP speakers were manually transcribed, 120 of these were analyzed by a speech therapist and 27 of them by four additional therapists. The tharapists marked 6 different processes/criteria like pharyngeal backing and hypernasality which typically occur in speech of people with CLP. We present detailed statistics about the the marked processes and the inter-rater agreement.

Keywords: Cleft lip and palate, pathologic speech, Children's speech

1. Introduction

Cleft Lip and Palate (CLP) is among the most frequent congenital abnormalities and has a birth prevalence ranging from 1/1000 to 2.69/1000 amongst different parts of the world (Mossey et al., 2009). The facial development is abnormal during gestation which leads to anatomic alterations with an insufficient closure of the lip, the palate and the jaw. Cleft lip and cleft palate can occur in combination or individually and can be present one sided (unilateral) or two sided (bilateral) (Godbersen, 1997), possibly including a gap in the jaw. Figure 1 shows examples of different cleft types: unilateral cleft lip, cleft palate, bilateral cleft lip and palate. These malformations may lead to various functional problems like disorders of hearing, swallowing and ingestion, breathing, and an affected articulation (Abramowicz et al., 2003). Due to the variety of CLP alterations the different phonemes are affected inhomogeneously for different patients.

A detailed phoneme analysis is needed in order to allow a speech therapy that fits the needs of an affected child and to allow a control of the therapy. In clinical routine, the perceptual analysis is done by expert listeners regarding different articulatory processes (Harding and Grunwell, 1998). Perceptual evaluations are subjective. Ratings of the same patients differ among raters, and are very time consuming; for each child about 3 hours are needed for the phoneme annotations. Thus, in clinical environment a strong demand for an objective, automatic analysis exists. However, perceptual evaluations are still the gold-standard in the clinical environment. Automatic systems have to be evaluated against perceptual evaluations. The reliability of an automatic system can be seen as sufficient when the human-machine-agreement is as high as the intra-rater-agreement.

For the development of an automatic system that gives an estimate on how heavily the different articulatory processes are affected, a large annotated corpus is necessary. We describe the dataset *Erlangen-CLP* and the perceptual annotations of the dataset. Results on inter-rater agreements are also discussed here.

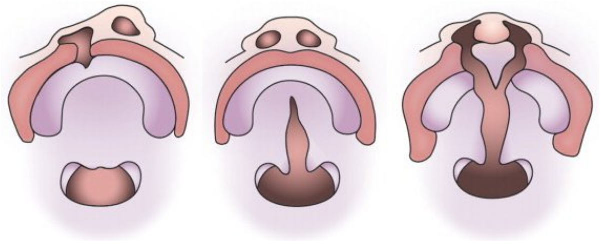


Figure 1: Examples of different cleft types: unilateral cleft lip (left), cleft palate (middle), bilateral cleft lip and palate (right) (Mossey et al., 2009).

The work deals with recordings of children speaking the PLAKSS (**P***sycho***l***inguistische* **A***nalyse* **K***indlicher* **S***prech*-**s***törungen*)-test (Fox, 2002), a semi-standardized test which is commonly used by speech therapists in German speaking countries. The test is composed of 99 pictograms (with 465 phonemes), which have to be named by the children. Three pictograms are shown on a single slide. The test contains all phonemes of the German language and the most important conjunctions among them at different word positions (beginning, central or ending). Figure 2 contains an example (the first slide of the test).

2. Speech Recordings

All children were recorded with the same microphone, a standard headset microphone (Plantronics Audio .655) with internal Analog-to-Digital-Converter in order to minimize the effects of varying recording equipment. We used PEAKS (Maier et al., 2009) to perform these recordings. Each slide of PLAKSS is shown on the computer screen. The child speaks the according words. Younger children try to name the pictograms, for older children the written

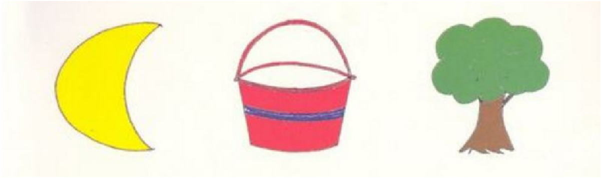


Figure 2: Pictograms of the first slide of the PLAKSS-test (Fox, 2002). The words are Mond (moon), Eimer (bucket), Baum (tree) focusing on phoneme /m/ at different word positions

words are presented. The naming of the pictograms is often problematic due to the (sometimes) ambiguous drawings. Trained students assist the children during speech recording and give them hints for finding the correct word without speaking it.

The naming of the pictograms induce two different problem: The use of word alternatives and the use of introductory words/sentences conjunctions, e.g., “this is a moon”, “... and this is a rabbit. My grandfather also has a rabbit. We gave him the name Schlachtreif, because we will eat him soon”. These problems complicate an automatic evaluation and require a reliable speech recognition engine. The interventions of the assisting students are also recorded during the sessions. This can also be problematic.

The recording of three pictograms in a row allows a rough segmentation and is processed by an speech recognition system. The collected data are transliterated manually by research assistant using a tool called Blitzscribe (Riedhammer, 2012) (see Figure 5). The transcripted data can be used to retrain or adapt the speech recognition engine and to evaluate performance on that. During transcription the research assistant marked the parts in the speech signal with interventions of the recording assistants, abruption of utterances and alternative words or synonyms.

An articulation assessment on word and/or phoneme level, requires a segmentation with time alignments indicating the time-boundaries of spoken words. Again, this can be automatically by speech recognition approaches and forced time-alignment. Due to the problems with the data described above, an automatic time alignment is prune to errors. Based on (manual or automatic) transcriptions, a speech recognition engine produces forced-alignments of the data. The research assistants verify this initial alignment and make corrections if necessary. A tool built with the visual components of the Java Speech Toolkit (JSTK) (Steidl et al., 2011) is used. A screenshot can be found in Figure 3. The manually segmented data can be used to measure the performance of the automatic segmentation approach. The performance of the automatic articulation assessment is either measured on automatically and manually segmented data.

Figure 4: Number of CLP speakers with respect to their age.

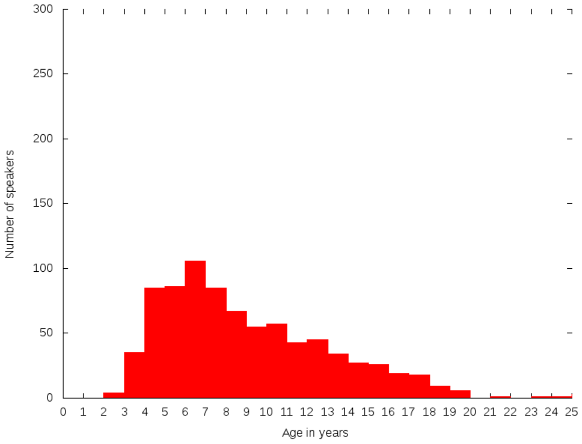
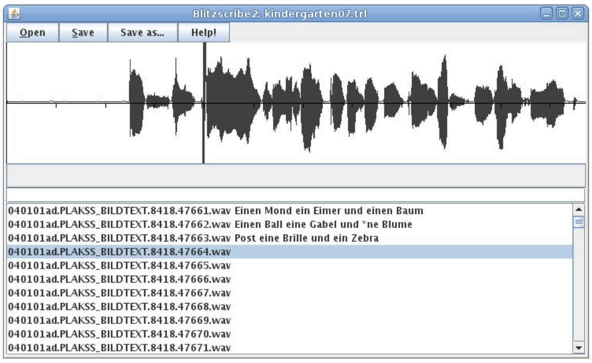


Figure 5: Blitzscribe: A tool for fast and efficient transcription tasks. See (Riedhammer, 2012) for details.



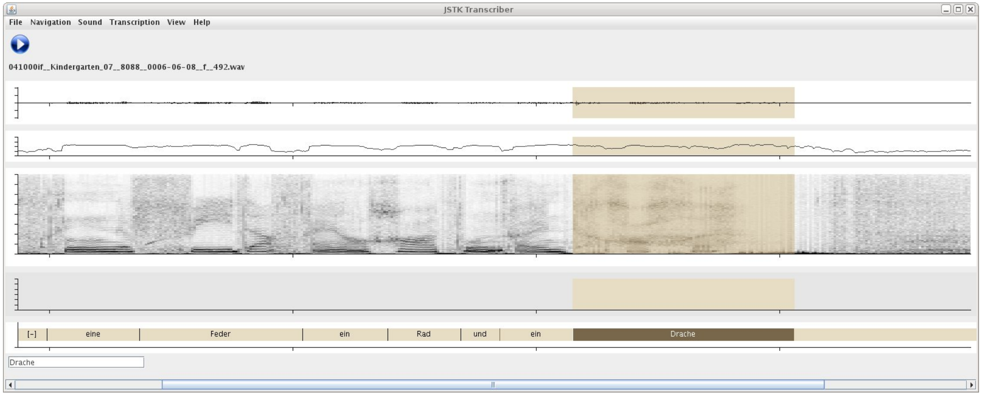
040101ad.PLAKSS_BILDTEXT.8418.47661.wav Einen Mond ein Eimer und einen Baum
040101ad.PLAKSS_BILDTEXT.8418.47662.wav Einen Ball eine Gabel und *ne Blume
040101ad.PLAKSS_BILDTEXT.8418.47663.wav Post eine Brille und ein Zebra
040101ad.PLAKSS_BILDTEXT.8418.47664.wav
040101ad.PLAKSS_BILDTEXT.8418.47665.wav
040101ad.PLAKSS_BILDTEXT.8418.47666.wav
040101ad.PLAKSS_BILDTEXT.8418.47667.wav
040101ad.PLAKSS_BILDTEXT.8418.47668.wav
040101ad.PLAKSS_BILDTEXT.8418.47669.wav
040101ad.PLAKSS_BILDTEXT.8418.47670.wav
040101ad.PLAKSS_BILDTEXT.8418.47671.wav

380 control speakers without CLP were recorded in pre- and primary schools in the region around Erlangen, Germany. 818 CLP speakers were recorded during routine examination in the University Clinic in Erlangen, Germany. The histogram of number of speakers vs. speaker age is shown in Figure 4. 355 speakers are female with the youngest speaker being 2 years old and the oldest speaker being 28. The mean age is 8.46 ± 12.6 years. 463 speakers are male. The youngest male speaker was 2 years old, the oldest one 30 at time of recording. The mean age for the male speakers is 8.85 ± 13.78 years. The presentation of pictograms allows children in preschool (who can not read) to denote the picture without letting them repeat the spoken words. However, this has the drawback of potential word alternatives. The assisting person gives hints in order to elicit the correct word. Therefor, all recordings were manually transcribed and automatically segmented on word level using a speech recognition system.

Out of the CLP-speakers we chose 250 speakers for a detailed processing. The speakers were selected so that we obtained a balanced set with respect to age, gender, and intelligibility, where the intelligibility was estimated using the word accuracy of a speech recognizer. The first part of Table 1 shows the statistics of the *control* corpus, the *clp* corpus, and the *clp-250* sub-corpus.

Among genders, the age distribution is roughly equal for *control* and *clp-250*. The automatic segmentation on word level was manually corrected and the speech of the assisting person was removed for *clp-250*. For the manual correction we used an open source tool called Blitzscribe (Riedhammer, 2012), which was developed in our group. In the following, we will concentrate on the 250 CLP and the 380

Figure 3: JSTKTrans: Tool for the manual correction of automatic alignments.



dataset	# female	# male	mean ± age
control	185	195	7.8 ± 10.4
clp	355	463	8.7 ± 13.3
clp-250	115	135	7.7 ± 9.5
clp-120	55	65	7.9 ± 7.8
clp-27	13	14	7.0 ± 6.2

Table 1: Number of speakers (male and female) and mean ± stddev statistics on the datasets. The second part of the table contains the statistics on the perceptually evaluated data. Clp-120 and clp-27 are subsets of clp.

control speakers.

3. Perceptual Annotations

Out of the clp corpus one speech therapist annotated 120 children regarding six different articulation processes. The processes are based on (Harding and Grunwell, 1998) and extended by (Wohlleben, 2004). They allow a phonetically-based differentiation of cleft palate and/or cleft lip speech. 27 children have been rated by four additional speech therapists. On average each speech therapist needed 3 hours to annotate a single child. During annotation, the speech therapists listened to each recording as often as they wanted to, and marked each conspicuous phoneme regarding one of the 6 processes/criteria:

Pharyngeal Backing (PB): The place of articulation is not correct. The tongue is shifted backward toward the pharynx during articulation.

Hypernasality (Hyper): The emission of air through the nose is excessive due to velopharyngeal insufficiency. This is very common in children with CLP.

Tension (Tens): The tension in articulation is diminished. This mostly results in a weakened pressure of consonants, e.g. a /p/ that is more articulated like a /b/.

Elision (Elis): A phoneme is not uttered and omitted. In CLP this is mostly due to a cleft in the palate.

Hyponasality (Hypo): The nasal emissions of air is missing. It makes the speaker sounds as if he has a cold.

Interdentality (Inter): Due to an improper closing of lip and jaw, the tip of the tongue becomes evident between upper and lower teeth.

In Table 2 the mean amount of marked phonemes (out of 465) per child is summarized for each rater. As an example: Rater1 marked 55.1 phonemes of 465 as Hypernasal on average per child with a standard deviation of 16.6. The table shows the marked phonemes with respect to the 6 different criteria on the clp-27 dataset. The row of criterion *all* denotes the mean amount of all marked phonemes per child. Please note, that the number is lower than the mean among the 6 criteria, since the raters sometimes marked one phone with different criteria, e.g., a phone can be pharyngeally backed and also be hypernasalized.

Hypernasality occurs most often, followed by hyponasality and tension. The number of marked phonemes differs largely between the different raters. Rater 5 marked much more phonemes than the other raters. This rater has the most experience in diagnosis and therapy of children with CLP. This rater also evaluated the clp-120 dataset.

In order to measure the inter-rater agreement among the five raters we performed pairwise inter-rater correlation experiments and calculated the average of them afterwards. Table 3 shows the average pair wise results of Spearman's correlation. E.g., column 1 shows the average pairwise correlations of rater1 with all other raters. We did not measure any significant differences between Pearson's and Spearman's correlation coefficient. The raters show a good interrater correlation for hypernasality ($\rho = 0.76$), pharyngeal backing ($\rho = 0.66$), interdentality ($\rho = 0.53$), and elision ($\rho = 0.52$). (Keuning et al., 1999) found similar values for perceptual ratings of hypernasality. Tension and hyponasality achieved a lower averaged pairwise correlation. This can be explained by the amount of marked phonemes in Table 2: For the criteria hypernasality, pharyngeal backing, and interdentality rater 1 to rater 4 marked a similar amount of phonemes. This is not the case for the criteria hyponasality and tension. It seems that these criteria are more difficult to rate. Rater 2 marked only 0.8 phonemes with the criterion tension and Rater 3 marked only 1.7 phonemes with hyponasality on average. There is a significant difference in the agreement of rater 2 to the other raters for the criterion tension. Rater 3 also showed a significant difference to the other raters for the criterion hyponasality.

4. Summary

The goal of our work is an automatic phoneme analysis of children with CLP in order to give an estimation on how strong different articulation processes are affected.

	rater1		rater2		rater3		rater4		rater5	
Crit	mean	stddev	mean	stddev	mean	stddev	mean	stddev	mean	stddev
Hyper	55.1	16.6	66.8	18.3	36.1	13.4	57.6	17.0	136.4	25.3
Hypo	9.1	6.8	62.9	17.8	1.7	2.9	20.0	10.0	52.2	15.7
Tens	1.7	2.9	0.8	2.0	8.8	6.6	104.6	22.9	93.7	21.0
Elis	5.1	5.0	4.4	4.7	1.2	2.5	5.6	5.3	28.0	11.5
PB	5.3	5.1	18.2	9.6	18.5	9.6	11.3	7.5	17.3	9.0
Inter	3.7	4.3	5.3	5.2	0.7	1.8	3.1	3.9	10.2	6.9
all	73.0	19.1	146.4	27.1	61.1	17.5	118.4	24.3	227.4	32.6

Table 2: Mean and standard deviation of number of marked phonemes of each rater in the 27 speaker dataset regarding the 6 criteria.

Crit	rater1	rater2	rater3	rater4	rater5	mean
Hyper	0.76	0.73	0.76	0.78	0.75	0.76
Hypo	0.41	0.37	0.27	0.38	0.50	0.39
Tens	0.40	0.27	0.45	0.41	0.44	0.39
Elis	0.50	0.60	0.47	0.61	0.44	0.52
PB	0.69	0.67	0.59	0.73	0.61	0.66
Inter	0.58	0.45	0.38	0.63	0.61	0.53
all	0.79	0.84	0.67	0.79	0.70	0.76

Table 3: Average pairwise inter-rater correlation regarding the 6 criteria

5. Acknowledgment

This work was funded by the German Research Foundation (DFG) under grant EY 15/18-2. We thank the speech therapists of our department for the perceptual evaluations.

6. References

Abramowicz, S., Cooper, M., Bardi, K., Weyant, R., and Marazita, M. (2003). Demographic and prenatal factors of patients with cleft lip and cleft palate. a pilot study. *J Am Dent Assoc*, 134(10):1371–6.

Fox, A. (2002). *PLAKSS - Psycholinguistische Analyse kindlicher Sprechstörungen*. Swets & Zeitlinger, Frankfurt a.M.i, Germany.

Godbersen, G. (1997). Das Kind mit Lippen-Kiefer-Gaumenspalte. *Laryngo-Rhino-Otologie*, 76:562–567.

Harding, A. and Grunwell, P. (1998). Active versus passive cleft-type speech characteristics. *International Journal of Language & Communication Disorders*, 33(3):329–352.

Keuning, K., Wieneke, G., and Dejonckere, P. (1999). The Intrajudge Reliability of the Perceptual Rating of Cleft Palate Speech Before and After Pharyngeal Flap Surgery: The Effect of Judges and Speech Samples. *Cleft Palate Craniofac J*, 36:328–333.

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., and Nöth, E. (2009). PEAKS - A system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5):425–437.

Mossey, P. A., Little, J., Munger, R. G., Dixon, M. J., and Shaw, W. C. (2009). Cleft lip and palate. *Lancet*, 374(9703):1773–1785, November.

Riedhammer, K. (2012). *Interactive Approaches to Video Lecture Assessment*. Ph.D. thesis, Technische Fakultät der Universität Erlangen–Nürnberg, Universität Erlangen–Nürnberg.

Steidl, S., Riedhammer, K., Bocklet, T., Hönig, F., and Nöth, E. (2011). Java Visual Speech Components for Rapid Application Development of GUI based Speech Processing Applications. In ISCA, editor, *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pages 3257–3260.

Wohlleben, U. (2004). *Die Verständlichkeitsentwicklung von Kindern mit Lippen-Kiefer-Gaumen-Segel-Spalten: Eine Längsschnittstudie über spalttypische Charakteristika und deren Veränderung*. Schulz-Kirchner-Verlag, Idstein, Germany.