AUTOMATIC PHONEME ANALYSIS IN CHILDREN WITH CLEFT LIP AND PALATE

Tobias Bocklet[1,2]*, Korbinian Riedhammer*[2]*, Ulrich Eysholdt*[1]*, Elmar Nöth*[2]

[1]Department of Phoniatrics and Pediatric Audiology, University Hospital Erlangen, Germany
[2]Pattern Recognition Lab, University of Erlangen-Nuremberg, Germany

ABSTRACT

Cleft Lip and Palate (CLP) is among the most frequent congenital abnormalities. The impaired facial development affects the articulation, with different phonemes being impacted inhomogeneously among different patients. This work focuses on automatic phoneme analysis of children with CLP for a detailed diagnosis and therapy control. In clinical routine, the state-of-the-art evaluation is based on perceptual evaluations. Perceptual ratings act as ground-truth throughout this work, with the goal to build an automatic system that is as reliable as humans. We propose two different automatic systems focusing on modeling the articulatory space of a speaker: one system models a speaker by a GMM, the other system employs a speech recognition system and estimates fMLLR matrices for each speaker. SVR is then used to predict the perceptual ratings. We show that the fMLLR-based system is able to achieve automatic phoneme evaluation results that are in the same range as perceptual inter-rater-agreements.

Index Terms— Pathology, automatic assessment, spectral features, GMM, fMLLR

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 [1]. 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) [2], 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 [3]. 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 [4]. 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 agreement between the automatic system is as high as the agreement among different human raters.

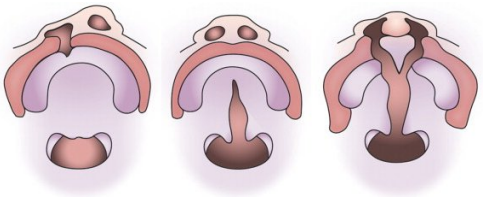

Fig. 1. Examples of different cleft types: unilateral cleft lip (left), cleft palate (middle), bilateral cleft lip and palate (right) [1].

The goal of this work is the development of an automatic system that gives an estimate on how strong the different articulatory processes are affected. The ground truth for our automatic system are speaker-wide scores that denote how many phonemes are affected with respect to different articulation processes. These scores have been estimated perceptually by clinical speech therapists. We propose two different automatic systems based on modeling the articulatory space of a speaker. The decision is on speaker level, rather than classification on phoneme level. This has the advantage, that no (automatic) phoneme segmentation is needed. In either system we start with Mel-Frequency Cepstrum Coefficients (MFCCs) as acoustic front-end. The first approach models the MFCCs with Gaussian Mixture Models (GMMs) by adapting an Universal Background Model (UBM) with Maximum A Posteriori (MAP) adaptation to the speaker-specific spectral features [5]. The mean-vectors of a speaker-specific GMM are then concatenated and used as GMM speaker-vector. In our second approach we use transformation matrices of Maximum Likelihood Linear Regression (MLLR) [6] adaptation as another kind of meta-features for acoustic speaker modeling. We use feature-space MLLR (fMLLR) [7] to produce a transformation matrix for each speaker. The elements of the matrix are appended and form an fMLLR speaker-vector. Both systems use a Support Vector Regression [8] in order to predict a speaker score.

The remainder of the paper is organized as follows. Section 2 shows the contributions of this work to current research. Section 3 describes the dataset and the perceptual annotations of the dataset. Results on inter-rater agreements are also discussed here. Section 4 introduces the two automatic systems. In Section 5, the results of the automatic systems are presented and discussed with respect to the inter-rater agreement results. The paper concludes with a summary and proposes future research.

2. CONTRIBUTIONS OF THIS WORK

In former work we focused on a holistic automatic evaluation of CLP speech where we tried to predict the level of intelligibility of 35 children with CLP. Best results were achieved with a speech recognition system [9] and acoustic modeling in form of GMM speaker-vectors [10, 11]. Most works in literature focus on single aspects like hyper-

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

Fig. 2. Pictograms of the first slide of the PLAKSS-test [15]. The words are Mond (moon), Eimer (bucket), Baum (tree) focusing on phoneme /m/ at different word positions

dataset	# female	# male	mean $\pm$ age
control	185	195	7.8 ± 10.4
clp	115	135	7.7 ± 9.5
clp-120	55	65	7.9 ± 7.8
clp-27	13	14	7.0 ± 6.2

Table 2. Number of speakers (male and female) and mean $\pm$ 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.

nasality on sustained vowels [12], not on spoken words/utterances. For a detailed analysis of CLP speech, more aspects have to be evaluated [13]. In [14] we focused on automatic detection of articulation disorders on a 26-speaker database. Perceptual annotation of one speech expert acted as ground truth. There we tackled the problem as a two-class task on frame, phoneme and word level. In a clinical point of view, a measurement on speaker-level is very important for the comparison of affected children and to allow therapy control. We introduce a new dataset of 380 healthy children and 250 children with CLP that contains phoneme annotated data of multiple raters. The annotations took 680 hours in total. In this work we use two different systems to achieve an automatic analysis on speaker level. The GMM-based speaker-vector system models the speaker without any prior knowledge. The fMLLR-based system trains a transformation matrix for each speaker and uses the manual word transcriptions to achieve that. The system's scenario is clearly the usage in clinical everyday life. The evaluation on real clinical data shows that the system is capable for this scenario.

3. DATA

The work deals with recordings of children speaking the PLAKSS (**P***sycho***l***inguistische* **A***nalyse* **K***indlicher* **S***prech***s***törungen*)-test [15], 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).

3.1. 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. 380 control speakers were recorded in pre- and primary schools in the region around Erlangen, Germany. 250 children with CLP were recorded during routine examination in the University Clinic in Erlangen, Germany. In either case, a person was assisting them. Each of the recordings was transliterated on word level. The first part of Table 2 shows the statistics of the *control* and *clp* corpus. Note that the control corpus and the clp corpus (without the clp-120 speakers) were used to train the UBM for the GMM speaker-vector system and the HMM for the fMLLR-speakervector system. Among genders, the age distribution is equal on both of the two sets.

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 allure the correct word. Thus, the data was manually segmented in order to get rid of the speech of the assisting person.

3.2. Perceptual Annotations

Out of the clp corpus one speech therapist annotated 120 children regarding six different articulation processes. The processes are based on [4] and extended by [16]. They allow a phonetically-based differentiation of cleft palate and/or cleft lip speech. 27 children have been rated by five 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.

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.

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

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

In Table 1 the mean amount of marked phonemes (out of 465) per child is summarized for each rater. 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 as 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 results of Spearman's correlation. We did not measure any significant differences between Pearson's and Spearman's correlation coefficient. The raters show a good inter-rater correlation for hypernasality ($\rho = 0.76$), pharyngeal backing ($\rho = 0.66$), interdentality ($\rho = 0.53$), and elision ($\rho = 0.52$). [17] 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 1: 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. AUTOMATIC SYSTEMS

The automatic systems model the acoustics of a speaker in two different ways. Both approaches use MFCCs as acoustic front-end with a frame rate of 10 ms and a frame size of 25 ms. For each frame t, the first 12 MFCCs and the log energy are retained together with their first and second order derivatives to construct a feature vector $\boldsymbol{x_t}$ with dimension $d = 39$. To minimize the influence of the microphone, Cepstral Mean Subtraction (CMS) is applied.

We use two different approaches to model the MFCCs either by GMMs in an unsupervised manner (see Section 4.1) or make use of manual word transcriptions and train speaker-specific fMMLR-transforms (see Section 4.2).

4.1. GMM speaker-vectors

This system is based on a statistical modeling of the articulatory space of a speaker. It relies on the assumption that the acoustics of pathologic speakers differ from those of healthy speakers. The degree of pathology is measured as the distance between the pathologic speaker model and a reference speaker model. The speaker model is a GMM representing all the MFCC vectors $\boldsymbol{x}$ being available for the speaker.

$$p(\boldsymbol{x}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \omega_i p_i(\boldsymbol{X}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}). \quad (1)$$

where ω_i, $\boldsymbol{\mu_i}$ and $\boldsymbol{\Sigma_i}$ denote the weights, the mean vectors and the covariance matrices of the different mixtures. A single, speaker-independent GMM is trained on the available data of healthy speakers, the so-called UBM. The Gaussians of this model are trained in an unsupervised manner by the *Expectation-Maximization* (EM) algorithm [18] in 10 iteration steps. The number of Gaussians is set to 128. An actual speaker model is derived by adapting the $\boldsymbol{\mu_i}$ of the UBM using the available speech of one speaker by *Maximum A Posteriori* (MAP) adaptation [18]. In order to reduce the dimensionality of the speaker model and to find a computationally more effective representation, only the mean vectors $\boldsymbol{\mu_i}$ are used to represent a speaker. They constitute a GMM speaker-vector. This vector then represents the acoustics properties of a certain speaker and has a dimension of 4992.

4.2. fMLLR speaker-vectors

This system is strongly motivated by the work of [19] and [20]. The basic idea of fMLLR is to transform the MFCCs in order to maximize the likelihood function given the MFCCs features and a Hidden Markov Model (HMM) by an affine transformation. The transformation consists of a square matrix A and a bias term b, reads as

$$\hat{x} = Ax + b \quad (2)$$

and can be rewritten as

$$\hat{\boldsymbol{x}} = \boldsymbol{W}\boldsymbol{x}^+, \text{where } \boldsymbol{x}^+ = \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix}. \quad (3)$$

$\boldsymbol{W}$ is a d by $d + 1$ matrix. The complex objective function consists of the log likelihood of the transformed MFCCs given the HMM-models, and the log determinant $log(|det\boldsymbol{A}|)$ and can be reformulated as a sum of $log(|det\boldsymbol{A}|)$ and a sum of quadratic functions of the rows of $\boldsymbol{W}$ [7].

We used the Kaldi-toolkit [21] for this automatic system. The initial speech recognition system was trained on the 380 healthy children and the 130 children with CLP that remain when the 120 perceptually evaluated speakers are excluded from the *clp* corpus. We trained the system with 2500 leaves and 15000 Gaussians in total.

We then aligned the speech data to the manual transcriptions and estimated speaker-specific fMLLR-transforms for each of the 120 perceptually evaluated speakers. The rows of the transform $\boldsymbol{W}$ are concatenated and form a 1560-dimensional fMLLR speaker-vector.

The difference to the GMM speaker-vector is the way of modeling: the GMM system accumulates the statistics in a completely unsupervised way, the fMLLR-system employs the models of a trained speech recognition system and uses aligned data to accumulate the statistics.

4.3. Prediction by Support Vector Regression

The prediction system is based on Support Vector Regression (SVR) [8]. The SVR is trained in a leave-one-speaker-out manner either us-

	GMM		fMLLR	
Crit	baseline	segmented	baseline	segmented
Hyper	0.57	0.60	0.81	0.82
Hypo	0.22	0.50	0.50	0.70
Tens	0.41	0.67	0.59	0.70
Elis	0.31	0.28	0.27	0.26
PB	0.56	0.58	0.47	0.59
Inter	0.57	0.60	0.72	0.74
all	0.52	0.62	0.72	0.79

Table 4. Results of the automatic Systems on the clp-27 subset. Target scores are the mean perceptual ratings of the 5 raters. The table shows the baseline results and the result achieved when segmenting and removing the assisting person during data recording.

	GMM		fMLLR	
Crit	baseline	segmented	baseline	segmented
Hyper	0.67	0.69	0.72	0.75
Hypo	0.29	0.27	0.53	0.55
Tens	0.55	0.57	0.53	0.52
Elis	0.39	0.43	0.39	0.44
PB	0.45	0.46	0.55	0.61
Inter	0.34	0.40	0.40	0.44
all	0.68	0.71	0.70	0.74

Table 5. Results of the automatic Systems on the clp-120 subset. Target scores are the perceptual ratings of *rater 5*. The table shows the baseline results and the result achieved when segmenting und removing the assisting person during data recording.

ing the GMM-based speaker-vectors or the fMLLR-based speaker-vectors as features. The targets for the SVR are the perceptual ratings of the speech therapists. On the clp-27 dataset ratings of five speech therapists exist for each of the six criteria. The mean value for each criterion provides the SVR targets on this dataset. The clp-120 data was rated by one rater only. The ratings of this rater provides the SVR targets on the clp-120 data.

5. RESULTS AND DISCUSSION

In Section 3.2 we already discussed the inter-rater correlation experiments. This section focuses on the results of the automatic systems. The results are divided into two subsections. Section 5.1 discusses the results on the clp-27 dataset. As ground-truth the average number of marked phonemes over the five raters was used. Section 5.2 focuses on the results on the clp-120 dataset. The ratings of one rater (em rater 5) acted as ground truth in this case.

For both sets of experiments we calculated the performance of the two systems for each of the 6 processes. Additionally, the sum of all processes was used as an overall score, i.e., number of affected phonemes per child. The two recognition systems were trained in either case on two different MFCCs sets: One set of features uses all the MFCCs frames of a child, this also might contain speech recordings of the person that assisted the child during speech recording. The assisting person was manually segmented and removed out of the speech signal in the other feature set.

5.1. Results on clp-27

Table 4 contains the result on the clp-27 subset. When focusing on the results of the GMM-based system it is clearly visible that segmenting the therapist makes a big difference for the criteria *hyponasality*, *tension* and the summation of affected phonemes (*sum*, last row). The distance of the assisting person and the unnatural direction to the microphone during recording affects the speech frames of the assistant. Recordings sound muffled and the pressure of plosives sounds unnaturally weakened and thus affect the results on the phoneme analysis of the child. The removal of the speech of the assistants lead to improved results. For the criteria *hyponasality*, *tension*, and *interdentality* the performance of the GMM-based system is in the same range as the inter-rater correlation results of the humans.

The improvement for the criteria *hyponasality*, *tension*, and *sum* is also valid for the fMLLR-based system. Compared to the GMM system the overall performance is improving for each criterion with the fMLLR system. The performance of the fMLLR-based system is in the same range as the inter-rater correlation results of the humans for each of the 7 criteria, except *elision*. However, this can be perfectly explained: The systems perform a time-invariant modeling of the speaker's articulation. Thus, omitted phonemes can not be modeled by these systems and the criterion *elision* can not be assessed properly.

5.2. Results on clp-120

Table 5 contains the result on the clp-120 subset. On this subset the improvement when removing the speech frames of the assisting persons is not as clear as on the clp-27 dataset. Small improvements are still visible among most of the criteria. The fMMLR-based system again outperforms the GMM-based system in almost all criteria. However, for the criterion *tension* the GMM-based system achieved a slightly higher correlation ($\rho_{GMM} = 0.57$ vs. $\rho_{fMLLR} = 0.53$) but the difference is not significant. Overall performance on the clp-120 set is lower than for the clp-27 set. We assume that is due to a more robust labeling when averaging over different raters. Comparing the clp-120 results with the inter-rater agreement of *rater 5* (second last column of Table 3) shows that the fMMLR-system is capable of modeling the articulation with "weak" labels of a single rater and achieve results that are in the same range as inter-rater agreement of this specific rater to the other raters.

6. SUMMARY

This work focused on automatic phoneme analysis of children with CLP. Our goal was to give an estimation on how strong different articulation processes are affected. We tried to predict a speaker wide score that reflects the number of affected phonemes regarding six different articulation processes; a summation of the affected phonemes acted as an overall score. We employed two different systems that perform a modeling of the articulatory space of a speaker. GMM speaker models were adapted from a UBM with one system, the other system modeled a speaker by fMLLR-transforms. Predictions were performed by SVR with either GMM speaker-vectors or fMLLR speaker-vectors as input vectors and perceptual labels of speech therapist as ground-truth scores. Experiments showed that the fMLLR-based system is capable of achieving automatic phoneme evaluation results that are in the same range as perceptual inter-rater-agreements. In future work we try to employ the idea of subspace Gaussian mixtures into the modeling approach and we focus on a more detailed phoneme analysis where we evaluate phonemes regarding their place of articulation.

7. REFERENCES

[1] P. A. Mossey, J. Little, R. G. Munger, M. J. Dixon, and W. C. Shaw, "Cleft lip and palate," *Lancet*, vol. 374, no. 9703, pp. 1773–1785, Nov. 2009.

[2] G. Godbersen, "Das Kind mit Lippen-Kiefer-Gaumenspalte," *Laryngo-Rhino-Otologie*, vol. 76, pp. 562–567, 1997.

[3] S. Abramowicz, M.E. Cooper, K. Bardi, R.J. Weyant, and M.L. Marazita, "Demographic and prenatal factors of patients with cleft lip and cleft palate. a pilot study.," *J Am Dent Assoc*, vol. 134, no. 10, pp. 1371–6, 2003.

[4] A. Harding and P. Grunwell, "Active versus passive cleft-type speech characteristics," *International Journal of Language & Communication Disorders*, vol. 33, no. 3, pp. 329–352, 1998.

[5] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, pp. 19–41, 2000.

[6] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[7] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[8] A.J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[9] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F Rosanowski, U Eysholdt, and E Nöth, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70/2006, pp. 1741–1747, 2006.

[10] T. Bocklet, A. Maier, K. Riedhammer, and E. Nöth, "Towards a Language-independent Intelligibility Assessment of Children with Cleft Lip and Palate," in *Proceedings of WOCCI 2009*, WOCCI, Ed., 2009, vol. 1, p. no pagination.

[11] T. Bocklet, K. Riedhammer, E. Nöth, U. Eysholdt, and T. Haderlein, "Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling," *Journal of Voice*, vol. 26, no. 3, pp. 390–397, 2012.

[12] S. Murillo Rendon, J.R. Orozco Arroyave, J.F. Vargas Bonilla, J.D. Arias Londono, and C.G. Castellanos Dominguez, "Automatic detection of hypernasality in children," in *New Challenges on Bioinspired Applications*. 2011, vol. 6687 of *Lecture Notes in Computer Science*, pp. 167–174, Springer Berlin Heidelberg.

[13] E. M. Konst, T. Rietveld, H. F. Peters, and H. Weersink-Braks, "Use of a perceptual evaluation instrument to assess the effects of infant orthopedics on the speech of toddlers with cleft lip and palate," *The Cleft Palate Craniofacial Journal*, vol. 40, no. 6, pp. 597–605, 2003.

[14] A. Maier, F. Hönig, T. Bocklet, E. Nöth, F. Stelzle, E. Nkenke, and M. Schuster, "Automatic detection of articulation disorders in children with cleft lip and palate," *Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2589–2602, 2009.

[15] A.V. Fox, *PLAKSS - Psycholinguistische Analyse kindlicher Sprechstörungen*, Swets & Zeitlinger, Frankfurt a.M., 2002.

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

[17] K. Keuning, G. Wieneke, and P. Dejonckere, "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*, vol. 36, pp. 328–333, 1999.

[18] J. L. Gauvain and C. H. Lee, "Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[19] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "Mllr transforms as features in speaker recognition," in *Proceedings of the* 6^{th} *Annual Conference of the International Speech Communication Association, Interspeech 2005*, 2005, pp. 2425–2428.

[20] M. Ferras, Cheung Chi Leung, C. Barras, and J.-L. Gauvain, "Constrained mllr for speaker recognition," in *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. IV 53– IV56.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. 2011, IEEE Signal Processing Society.