Automatic Evaluation of Characteristic Speech Disorders in Children with Cleft Lip and Palate

Andreas Maier[1,2], *Florian Hönig*[1], *Christian Hacker*[3], *Maria Schuster*[2], *Elmar Nöth*[1]

[1] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany
[2]Abteilung für Phoniatrie und Pädaudiologie, Universitätsklinikum Erlangen, Germany
[3]Elektrobit Automotive GmbH, Erlangen, Germany
Andreas.Maier@cs.fau.de

Abstract

This paper discusses the automatic evaluation of speech of children with cleft lip and palate (CLP). CLP speech shows special characteristics such as hypernasality, backing, and weakening of plosives. In total five criteria were subjectively assessed by an experienced speech expert on the phone level. This subjective evaluation was used as a gold standard to train a classification system. The automatic system achieves recognition results on frame, phone, and word level of up to 75.8 % CL. On speaker level significant and high correlations between the subjective evaluation and the automatic system of up to 0.89 are obtained.

Index Terms: pathologic speech, speech assessment, pronunciation scoring, children's speech

1. Introduction

Cleft Lip and Palate (CLP) is the most common malformation of the head. It constitutes almost two-thirds of the major facial defects and almost 80 % of all orofacial clefts [1]. Its prevalence differs in different populations from 1 in 400 to 500 newborns in Asians to 1 in 1500 to 2000 in African Americans. The prevalence in Caucasians is 1 in 750 to 900 births [2, 3].

In clinical practice, articulation disorders are mainly evaluated by subjective tools. The simplest method is the auditive perception, mostly performed by a speech therapist. Previous studies have shown that experience is an important factor that influences the subjective estimation of speech disorders which leads to inaccurate evaluation by persons with only few years of experience as speech therapist [4]. Until now, objective means exist only for quantitative measurements of nasal emissions [5, 6, 7] and for the detection of secondary voice disorders [8]. But other specific articulation disorders in CLP cannot be sufficiently quantified.

In this paper, we present a new technical procedure for the measurement and evaluation of specific speech disorders and compare the results obtained with subjective ratings of an experienced speech therapist.

2. Speech of Children with Cleft Lip and Palate

The effects of CLP on the speech of children are manifold. The most important aspects can be formulated as:

- Significant differences between normal children and CLP children were measured with the subjective assessment of the intelligibility and the measurement of the nasal airflow [9, 6, 10]. The speech exhibits hypernasality (HN) in vowels (perceived as characteristic "nasality") and nasalized consonants (NC) with characteristic noise in high frequencies.
- Speech of CLP children contains typical cleft type characteristics — pharyngeal backing (PB), glottal articulation (also called laryngeal replacement, LR), and absent or weakened pressure consonants (WP) [11].
- No significant differences between isolated cleft lip and CLP exist in the frequency of occurrence of certain speech disorders. Furthermore, the speech outcome is similar in cleft palate and CLP children [12]. Therefore, CLP is not further differentiated in this study.

3. CLP Speech Data

A group of 26 children (5 female and 21 male) was recorded using a standard head set (dnt Call 4U Comfort). The children spoke the PLAKSS Test[1] [13] a German semi-standardized test which is commonly used by speech therapists. Two of the children in the dataset had an isolated cleft lip, three an isolated cleft palate, 19 unilateral CLP and another two bilateral CLP.

The speech data were manually transliterated. Using an automatic speech recognition system the data were segmented into 7647 phones and 1916 words. The data were sampled at 16 kHz with a quantization of 16 bit. The dataset is a subset of the data which has already been investigated in [14].

4. Subjective Evaluation of the Speech Data

The speech therapist had been working with children with cleft lip and palate for many years. Therefore, she could differentiate all criteria as listed in Table 1. The subjective phone level evaluation was very time-consuming: The evaluation of the speech data of a single child (about three to four minutes of speech data) took about one hour.

Table 1 lists the result of the phone level evaluation. All of the articulation errors are sparse in the data set. The table presents only the number of misarticulated phones. The number of wrongly pronounced words is almost the same since a single articulation error within a word was sufficient to count the whole word as disordered. Only two words in the dataset contained two times the same type of articulation error, i.e., 33 words with PB and 31 word with LR were annotated. The last column shows the number of children which were affected by the different disorders. While PB, LR, and HN appear in only few children WP and NC appear in more than half of the children.

[1] "Psycho-Linguistic Analysis of Speech Disorders" ("Psycho-Linguistische Analyse kindlicher Sprechstörungen" in German)

Table 1: Serious articulation errors were annotated in the data by an experienced speech therapist according to the taxonomy of [11] in the group of 26 CLP children

Error	Description	Abbr.	Occ.	# of Affected Children
hypernasality in vowels	the nasal air flow is persistent throughout the vowel	HN	49	4
nasalized consonants	the consonants are nasalized i.e., air is emitted through the nose during the articulation of the consonants	NC	329	15
pharyngealization	tongue is shifted backwards towards the pharynx during articulation	PB	34	7
glottal articulation	the closure of the plosives is done in a glottal manner instead of a labial. The disorder is also called laryngeal replacement in the following.	LR	32	4
weakened pressure consonants	plosives are not formed or weakened during the articulation	WP	105	14

In 7 children less than 15 % of the words were marked as disordered and in 3 children not a single word was affected. This was to be expected since some of the children have normal or almost normal speech. The prior distribution of the classes was not changed for the classification task, since we wanted to keep the experiments as realistic as possible.

5. Automatic Evaluation System

Figure 1 shows the experimental setup. As a typical classification system [15], it is divided into the blocks preprocessing, feature extraction, classification, and results in a decision for a class. The procedure is performed on frame, phone, word, and speaker level. From the respective result of the lower level, meta features are computed and supplied to the respective higher level. As meta features the mean, the maximum, the minimum, the standard deviation, the sum, and the product of the output probabilities are computed. Furthermore, we also regard the absolute and relative frequency of classes as meta features. In the following the classification system is described.

5.1. Preprocessing

As already mentioned the preprocessing is currently performed semi-automatically. The speech data was transliterated by hand. In the future, we plan to replace this step by a fully automatic procedure. Next, the transliteration is aligned by a speech recognition engine as described by Stemmer in [16]. This procedure yields estimated positions of words and phonemes in the signal which is used to segment the audio data accordingly before feature extraction.

5.2. Feature Extraction

In our classification system state-of-the-art features for the evaluation of speech are employed. Previously, good correlations between the intelligibility and the recognition accuracy (*RecAcc*) were reported. Furthermore, the visualization with Sammon's mapping also yields a representation which is connected to the intelligibility (*2-D Sammon Coordinates* and *3-D Sammon Coordinates*). On word level prosodic (*ProsFeat*) and pronunciation (*PronFexW*) features have been shown to be useful for the assessment of speech data. Certain pronunciation features are already available on phone level (*PronFexP*). The Teager Energy Profile (*TEP*) is a well known feature for the detection of hypernasality in vowels. The Teager Energy operator (TEO) is defined as:

$$\psi[f(n)] = [f(n)]^2 - f(n+1)f(n-1) \qquad (1)$$

$f(n)$ denotes the time domain audio signal. The TEO's output is called the *TEP*. One frame level Mel Frequency Cepstrum Coefficients (*MFCCs*) are well known to hold relevant information for the articulation. Table 2 lists all used features and references to further literature.

5.3. Classification

For the classification various classifiers as provided in the WEKA toolbox [23] were employed. The following classifiers were employed:

- **OneR**: An interval-based classifier
- **DecisionStump**: A threshold-based classifier
- **LDA-Classifier**: Classification based on Linear Discriminant Analysis (LDA)
- **NaiveBayes**: Classification according to a unimodal Gaussian distribution
- **J48**: A C4.5 Decision Tree
- **PART**: The PART creates partial C4.5 trees in each iteration and creates a rule from the "best" leaf.
- **RandomForest**: A classifier built from many random trees
- **SVM**: Support Vector Machines
- **AdaBoost**: A boosted version of any of the above classifiers

Each of the classifiers was tested on each level. The use of different classifiers on different levels was also allowed. Use of different classifiers on the same level, however, was not permitted, e.g. different classifiers for different phonemes.

6. Experimental Results

All experiments on frame, phone, and word level were conducted as leave-one-speaker-out evaluation. As measure for the accuracy the class-wise averaged recognition rate (CL), i.e., the unweighted average recall, and the absolute recognition rate (RR) are reported. The recall is defined as the number of true positives divided by the number of true positives and false negatives and is, therewith, equal to the definition of the sensitivity. In order to optimize the CL, the training samples were weighted to form a balanced training set.

As reported in Table 3 very high values are reached for RR. This, however, is related to the unbalanced test sets: Most samples in the test set are not pathologic. Hence, classification of all samples to the class "normal" already yields high RRs. The CL shows that the accuracy is moderate in most cases for these two class problems.

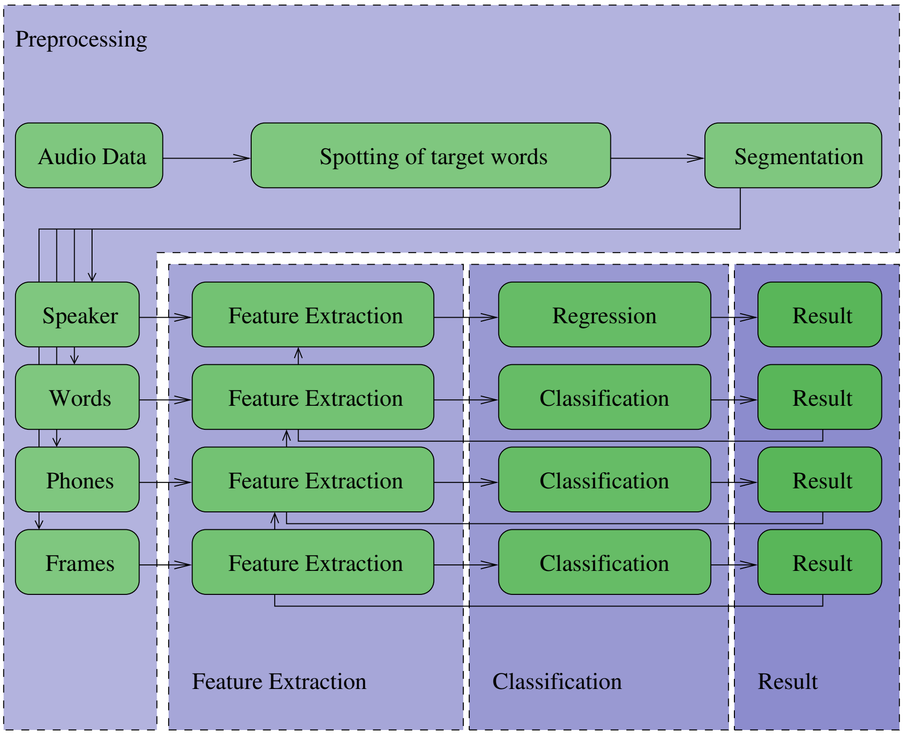

Figure 1: Experimental setup of the classification system: Right after the recording the preprocessing is performed. The data are transliterated manually and segmented automatically. Next, the feature extraction takes place on each level. Then, the features are supplied to a classifier which performs evaluation.

For each criterion the detected percentage of pathologic words was correlated with the subjectively marked percentage using Pearson's correlation [24] to evaluate the system on speaker level. Furthermore, the speaker level features *RecAcc*, *2-D Sammon Coordinates*, and *3-D Sammon Coordinates* were added to the regression. Significance tests revealed that all reported correlations are significant with $p < 0.01$.

7. Discussion

This paper presented the first semi-automatic evaluation system for the characteristic speech disorders of children with CLP. On frame and phone level CLs of up to 71.1 % were reached. On word level the best CL was 75.8 %. This is comparable to other studies concerning pronunciation scoring [25, 26, 19].

The classification errors seem to be systematic, because the classification on word level with up to 75.8 % CL is already sufficient for a good quantification of all five disorders on speaker level. The lowest correlations was found with 0.70 for pharyngeal backing (PB) while the best correlation was 0.89 for hypernasality. All correlations were significant with $p < 0.01$.

A correlation of 0.80 is already reliable enough to quantify speech disorders on speaker level sufficiently. This is in the same range as human raters would agree with each other [4] for this task. In the future the integration of an automatic speech recognition system will replace the manual transliteration.

8. Acknowledgments

This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SCHU2320/1-1. We would like to thank Dr. Ulrike Wohlleben for the expert's annotation of the data.

9. References

[1] B. Eppley, J. van Aalst, A. Robey, R. Havlik, and M. Sadove, “The spectrum of orofacial clefting,” *Plastic and Reconstructive Surgery*, vol. 115, no. 7, pp. 101–114, 2005.

[2] M. Tolarova and J. Cervenka, “Classification and birth prevalence of orofacial clefts,” *Am J Med Genet*, vol. 75, no. 2, pp. 126–137, 1998.

[3] H. Kawamoto, “Rare craniofacial clefts,” in *Plastic Surgery*, J. C. McCarthy, Ed., vol. 4. USA: Saunders, Philadelphia, 1990.

[4] S. Paal, U. Reulbach, K. Strobel-Schwarthoff, E. Nkenke, and M. Schuster, “Beurteilung von Sprechauffälligkeiten bei Kindern mit Lippen-Kiefer-Gaumen-Spaltbildungen,” *J Orofac Orthop*, vol. 66, no. 4, pp. 270–278, 2005.

[5] C. Küttner, R. Schönweiler, B. Seeberger, R. Dempf, J. Lisson, and M. Ptok, “Objektive Messung der Nasalanz in der deutschen Hochlautung,” *HNO*, vol. 51, pp. 151–156, 2003.

[6] K. V. Lierde, M. D. Bodt, J. V. Borsel, F. Wuyts, and P. V. Cauwenberge, “Effect of cleft type on overall speech intelligibility and resonance,” *Folia Phoniatrica et Logopaedica*, vol. 54, no. 3, pp. 158–168, 2002.

[7] T. Hogen Esch and P. Dejonckere, “Objectivating Nasality in Healthy and Velopharyngeal Insufficient Children with the Nasalance Acquisition System (NasalView): Defining Minimal Required Speech Tasks Assessing Normative Values for Dutch Language,” *Int J Pediatr Otorhinolaryngol*, vol. 68, no. 8, pp. 1039–46, 2004.

[8] T. Bressmann, R. Sader, M. Merk, W. Ziegler, R. Busch, H. Zeilhofer, and H. Horch, “Perzeptive und apparative Untersuchung der Stimmqualität bei Patienten mit Lippen-Kiefer-Gaumenspalten,” *Laryngorhinootologie*, vol. 77, no. 12, pp. 700–708, 1998.

[9] J. Karling, O. Larson, R. Leanderson, and G. Henningsson, “Speech in Unilateral and Bilateral Cleft Palate Patients from

Table 2: Overview on the feature sets which are extracted on the four different evaluation levels

Label	Level	#	Description	Reference
RecAcc	speaker	2	Accuracy of the speech recognition (word correctness and accuracy)	[14]
2-D Sammon Coordinates	speaker	2	Coordinates on a 2-D Sammon map	[17]
3-D Sammon Coordinates	speaker	2	Coordinates on a 3-D Sammon map	[17]
ProsFeat	word	37	Features based on the energy, the F_0, pauses, and duration to model the prosody of the speaker	[18]
PronFexW	word	7	Pronunciation features (PronFex) to score the correctness of the current word	[19]
PronFexP	phone	6	Features to score the correctness of the Pronunciation (PronFex) of the current phone	[20]
TEP	phone	1	Teager Energy Profile to detect nasality in vowels	[21]
MFCCs	frame	24	Mel Frequency Cepstrum Coefficients	[22]

Table 3: Overview on the results of the pronunciation assessment on frame, phone, word, and speaker level: All reported correlations are significant at $p < 0.01$.

	Automatic Evaluation						
	Frame		Phone		Word		Speaker
Criterion	CL	RR	CL	RR	CL	RR	r
HN	56.8 %	99.0 %	62.9 %	99.0 %	60.6 %	96.9 %	0.89
NC	62.0 %	94.2 %	68.5 %	95.6 %	63.6 %	82.5 %	0.85
PB	66.0 %	99.1 %	76.9 %	99.6 %	67.9 %	98.2 %	0.70
LR	59.8 %	99.6 %	69.5 %	99.6 %	63.8 %	98.2 %	0.81
WP	71.1 %	97.8 %	71.1 %	97.8 %	75.8 %	97.8 %	0.82

Stockholm," *Cleft Palate-Craniofacial Journal*, vol. 30, no. 1, pp. 73–77, 1993.

[10] K. V. Lierde, M. D. Bodt, I. Baetens, V. Schrauwen, and P. V. Cauwenberge, "Outcome of Treatment regarding Articulation, Resonance and Voice in Flemish Adults with Unilateral and Bilateral Cleft Palate," *Folia Phoniatrica et Logopaedica*, vol. 55, pp. 80–90, 2003.

[11] D. Sell, P. Grunwell, S. Mildinhall, T. Murphy, T. Cornish, D. Bearn, W. Shaw, J. Murray, A. Williams, and J. Sandy, "Cleft Lip and Palate Care in the United Kingdom—The Clinical Standards Advisory Group (CSAG) Study. Part 3: Speech Outcomes," *Cleft Palate-Craniofacial Journal*, vol. 32, no. 1, pp. 30–37, 2001.

[12] M. Timmons, R. Wyatt, and T. Murphy, "Speech after repair of isolated cleft palate and cleft lip and palate," *Britisch Journal of Plastic Surgery*, vol. 54, pp. 377–384, 2001.

[13] A. Fox, "PLAKSS – Psycholinguistische Analyse kindlicher Sprechstörungen," Swets & Zeitlinger, Frankfurt a.M., Germany, now available from Harcourt Test Services GmbH, Germany, 2002.

[14] A. Maier, C. Hacker, E. Nöth, E. Nkenke, T. Haderlein, F. Rosanowski, and M. Schuster, "Intelligibility of children with cleft lip and palate: Evaluation by speech recognition techniques," in *Proc. International Conf. on Pattern Recognition (ICPR)*, vol. 4, Hong Kong, China, 2006, pp. 274–277.

[15] H. Niemann, *Klassifikation von Mustern*, 2nd ed. available online, 2003, http://www5.informatik.uni-erlangen.de/Personen/niemann/klassifikation-von-mustern/m00links.html; last visited 02/12/2008.

[16] G. Stemmer, *Modeling Variability in Speech Recognition*. Berlin, Germany: Logos Verlag, 2005.

[17] T. Haderlein, D. Zorn, S. Steidl, E. Nöth, M. Shozakai, and M. Schuster, "Visualization of Voice Disorders Using the Sammon Transform," in *9th International Conf. on Text, Speech and Dialogue (TSD)*, ser. Lecture Notes in Artificial Intelligence, P. Sojka, I. Kopeček, and K. Pala, Eds., vol. 4188. Berlin, Heidelberg, New York: Springer, 2006, pp. 589–596.

[18] T. Haderlein, E. Nöth, M. Schuster, U. Eysholdt, and F. Rosanowski, "Evaluation of Tracheoesophageal Substitute Voices Using Prosodic Features," in *Proc. Speech Prosody, 3rd International Conference*, R. Hoffmann and H. Mixdorff, Eds. Dresden, Germany: TUDpress, 2006, pp. 701–704.

[19] C. Hacker, T. Cincarek, A. Maier, A. Heßler, and E. Nöth, "Boosting of Prosodic and Pronunciation Features to Detect Mispronunciations of Non-Native Children," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. Hawaii, USA: IEEE Computer Society Press, 2007, pp. 197–200.

[20] T. Cincarek, "Pronunciation Scoring for Non-Native Speech," Diplomarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 2004.

[21] D. Cairns, J. Hansen, and J. Riski, "A Noninvasive Technique for Detecting Hypernasal Speech using a nonlinear Operator," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 1, pp. 35–45, 1996.

[22] S. Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing (ASSP)*, vol. 28, no. 4, pp. 357–366, 1980.

[23] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Fransisco, CA, USA: Morgan Kaufmann, 2005.

[24] K. Pearson, "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia," *Philosophical Transactions of the Royal Society of London*, vol. 187, pp. 253–318, 1896.

[25] A. Neri, C. Cuchiarini, and C. Strik, "Feedback in Computer Assisted Pronunciation Training: Technology Push or Demand Pull?" in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Orlando, USA: IEEE Computer Society Press, 2002, pp. 1209–1212.

[26] C. Hacker, T. Cincarek, R. Gruhn, S. Steidl, E. Nöth, and H. Niemann, "Pronunciation feature extraction," in *Pattern Recognition, 27th DAGM Symposium, August 30 - September 2005, Vienna, Austria, Proceedings*, ser. Lecture Notes in Computer Science, G. Kropatsch, R. Sablatnig, and A. Hanbury, Eds., vol. 3663. Berlin, Heidelberg,Germany: Springer, 2005, pp. 141–148.