Automatic Assessment of Children's Speech with Cleft Lip and Palate

Andreas Maier[‡*], Elmar Nöth[*], Emeka Nkenke[†], Maria Schuster[‡]

[*] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg
Martensstraße 3, 91058 Erlangen, Germany
Andreas.Maier@cs.fau.de

[†]Mund-, Kiefer- und Gesichtschirurgische Klinik, Universität Erlangen-Nürnberg,
Glückstraße 11, 91054 Erlangen, Germany

[‡]Abteilung für Phoniatrie und Pädaudiologie, Universität Erlangen-Nürnberg
Bohlenplatz 21, 91054 Erlangen, Germany

Abstract

Cleft lip and palate (CLP) may cause functional limitations even after adequate surgical and non-surgical treatment, speech disorder being one of them. Until now, an automatic, objective means to determine and quantify the intelligibility did not exist. We have created an automatic evaluation system that assesses speech, based on the result of an automatic speech recognizer. It was applied to 35 recordings of children with CLP. A subjective evaluation of the intelligibility was performed by two experts and confronted to the automatic speech evaluation. It complied with experts' rating of intelligibility. Furthermore we present the results obtained on a control group of 45 recordings of normal children and compare these results with those of the CLP children.

Samodejna ocena govora otrok z zajčjo ustnico in volčjim žrelom

Zajčja ustnica in volčje žrelo lahko povzročata funkcijske omejitve tudi po ustreznem operativnem ali neoperativnem zdravljenju, med njimi so tudi motnje govora. Do sedaj ni obstajal samodejni objektivni način ugotavljanja razumljivosti. Razvili smo sistem za samodejno vrednotenje, ki ocenjuje govor na podlagi rezultatov samodejnega razpoznavalnika govora. Uporabljen je bil pri 35 posnetkih otrok z zajčjo ustnico in volčjim žrelom. Subjektivno vrednotenje razumljivosti, ki sta ga opravila dva strokovnjaka, je bilo soočeno s samodejnim vrednotenjem govora. Slednje se je ujemalo z oceno razumljivosti strokovnjakov. Poleg tega predstavljamo rezultate, pridobljene pri kontrolni skupini s 45 posnetki govora otrok brez motenj govora, in jih primerjamo z rezultati posnetkov govora otrok z zajčjo ustnico in volčjim žrelom.

1. Introduction

Cleft lip and palate (CLP) is the most common malformation of the head. It can result in morphological and functional disorders (Wantia and Rettinger, 2002), whereat one has to differentiate primary from secondary disorders (Millard and Richman, 2001; Rosanowski and Eysholdt, 2002). Primary disorders include e.g. swallowing, breathing and mimic disorders. Speech and voice disorders (Schönweiler and Schönweiler, 1994) as well as conductive hearing loss that affect speech development (Schönweiler et al., 1999), are secondary disorders. Speech disorders can still be present after reconstructive surgical treatment. The characteristics of speech disorders are mainly a combination of different articulatory features, e.g. enhanced nasal air emissions that lead to altered nasality, a shift in localization of articulation (e.g. using a /d/ built with the tip of the tongue instead of a /g/ built with back of the tongue or vice versa), and a modified articulatory tension (e.g. weakening of the plosives /t/, /k/, /p/) (Harding and Grunwell, 1998). They affect not only the intelligibility but therewith the social competence and emotional development of a child. 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 leading to inaccurate evaluation by persons with only few years of experience (Paal et al., 2005). Until now, objective means exist only for quantitative measurements of nasal emissions (Küttner et al., 2003; Lierde et al., 2002; Hogen Esch and Dejonckere, 2004) and for the detection of secondary voice disorders (Bressmann et al., 1998). But other specific or non-specific articulation disorders in CLP as well as a global assessment of speech quality cannot be sufficiently quantified. In this paper, we present a new technical procedure for the measurement and evaluation of speech disorders and compare the results obtained with subjective ratings of a panel of expert listeners.

2. Automatic Speech Recognition System

For the objective measurement of the intelligibility of children with speech disorders, an automatic speech recognition system was applied, a state-of-the-art word recognition system developed at the Chair for Pattern Recognition (Lehrstuhl für Mustererkennung) of the University of Erlangen. In this study, the latest version as described in detail in (Stemmer, 2005) was used. The recognizer can handle spontaneous speech with mid-sized vocabularies of up to 10,000 words. As features we use Mel-Frequency Cepstrum Coefficients (MFCC) 1 to 11 plus the energy of the signal. Additionally 12 delta coefficients are computed over a context of 2 time frames to the left and the right side (56 ms in total). The recognition is performed with semi-continuous Hidden Markov Models (SCHMMs). The codebook contains 500 full covariance Gaussian densities which

are shared by all HMM states. The elementary recognition units are polyphones (Schukat–Talamazzini and Niemann, 1991). The polyphones were constructed for each sequence of phones which appeared more than 50 times in the training set.

We used two types of unigram language models according to the application scenario. This helps to enhance recognition results by including linguistic information. However, for our purpose it was necessary to put more weight on the recognition of acoustic features. In the first scenario the transliteration is assumed to be unknown. So we created a basic language model which was trained with just the reference words of the test (see below) since no further information was available. This model has a perplexity of 43 on the reference text. In the second scenario the transliteration is available. This means that the data have to be transliterated completely and that additional words can appear which were not in the set of the reference words. These words are added to the language model in order to enable the their recognition. However, the probability of the target words is increased by a factor of 2. The test set perplexity of the language model differs for each speaker since the language model is constructed individually if the transliteration is known.

The speech recognition system had been trained with acoustic information from spontaneous dialogues of the VERBMOBIL project (Wahlster, 2000) and normal children's speech. The speech data of non-pathologic children voices (30 female and 23 male) were recorded at two local schools (age 10 to 14) in Erlangen and consisted of read texts. The training population of the VERBMOBIL project consisted of normal adult speakers from all over Germany and thus covered all dialectal regions. All speakers were asked to speak "standard" German. 90 % of the training population (47 female and 85 male) were younger than 40 years. During training an evaluation set was used that only contained children's speech. The adults' data was adapted by vocal tract length normalization as proposed in (Stemmer et al., 2003).

MLLR adaptation (Gales et al., 1996) with the patients' data lead to further improvement of the speech recognition system.

3. Data

All children were asked to name pictures that were shown according to the PLAKSS test (Fox, 2002). This German test consists of 99 words shown as pictograms on 33 slides. With this test, the speech of children can be evaluated even if they are quite young since they do not need the ability to read. However, the children could take advantage of being able to read since the reference words were shown as subtitles. The test includes all possible phonemes of the German language in different positions (beginning, center and end of a word).

The patients' group consisted of 35 children and adolescents (13 girls and 22 boys) with CLP at the age from 3.3 to 18.5 years (mean 8.3 $\pm$ 3.6 years). The examination was included in the regular out-patient examination of all children and adolescents with CLP. These speech samples were recorded with a close-talking microphone (dnt Call 4U Comfort headset) at a sampling frequency of 16 kHz and quantized with 16 bit. For these data no further post-processing was done.

Furthermore a control group with 45 normal children was recorded at a local elementary school. In total, data from 27 girls and 18 boys were collected. The children were in the age from 7.4 to 10.7 (mean 9.5 $\pm$ 0.9 years). The data were collected at 48 kHz with 16 bit quantization. To match the patients' data a resampling to 16 kHz was done. For the control group a Sennheiser close-talking microphone (handgrip K3U with ME 80 head) was used. These data were post-processed: In some cases the voice of the instructor was audible on the sound track. So the instructor's voice was removed in all occasions. Furthermore all of the children's speech data was transliterated.

Informed consent had been obtained by all parents of the children prior to the recording. All children were native German speakers, some using a local dialect.

4. Subjective Evaluation

Two voice professionals subjectively estimated the intelligibility of the children's speech while listening to a play-back of the recordings. A five point Likert scale (1 = very high, 2 = rather high, 3 = medium, 4 = rather low, 5 = very low) was applied to rate the intelligibility of all individual turns. In this manner an averaged mark – expressed as a floating point value – for each patient could be calculated.

5. Analysis and Automatic Evaluation

For the agreement computations between different raters on the one hand and raters/recognizer on the other hand we use the Pearson product-moment correlation coefficient (Pearson, 1896). It allows to compare two number series which are of different scale and margin like in the given case. So the ratings of the human experts and those of the speech recognition system can be compared directly without having to define a mapping between word accuracies and Likert scores. In order to compare both raters to the recognition system the average rating of the experts was computed for each speaker. For the recognition rate of the speech recognition system we investigated the word accuracy (WA) like in (Haderlein et al., 2004), (Schuster et al., 2005) or (Maier et al., 2006; Schuster et al., 2006) and the word recognition rate (WR). The WA is defined as

$$WA = \frac{C - I}{R} \cdot 100\,\%$$

where C is the number of correctly recognized words, I the number of wrongly inserted words and R the number of words in the reference text. The WR is defined as follows:

$$WR = \frac{C}{R} \cdot 100\,\%$$

Both measurements need a reference text in order to determine the number of correctly recognized words. However, since the reference are pictures, the text is not known a priori. One solution to this problem is to transliterate all the data like it was done before. Since we developed a new recording and evaluation software we now know the exact time when the reference slide was moved to the next slide.

measurement	recognized word chain	reference	%
transliteration WA	This is moon, bucket and a a ball	This is a moon, a bucket, and a tree	55.5
transliteration WR	This is moon, bucket and a a ball	This is a moon, a bucket, and a tree	66.6
automatic WA	tiger moon bucket apple ball	moon bucket tree	0
automatic WR	tiger moon bucket apple ball	moon bucket tree	66.6

Table 1: Example of the effects of the automatic reference on the WA and WR. We assume that the spoken utterance is "This is a moon, a bucket, and a tree". Thus, the automatic reference is "moon bucket tree"

measurement	transliteration WA	transliteration WR
automatic WA	0.40	0.21
automatic WR	0.60	0.60

Table 2: Correlation between the different measurements regarding the control group. The automatic WR yields the results with the best correlation to the transliteration-based measurements

rater	M	S	mean
automatic WA	-0.83	-0.77	-0.82
automatic WR	-0.88	-0.85	-0.89

Table 3: Correlation between the different raters and the automatic measurements

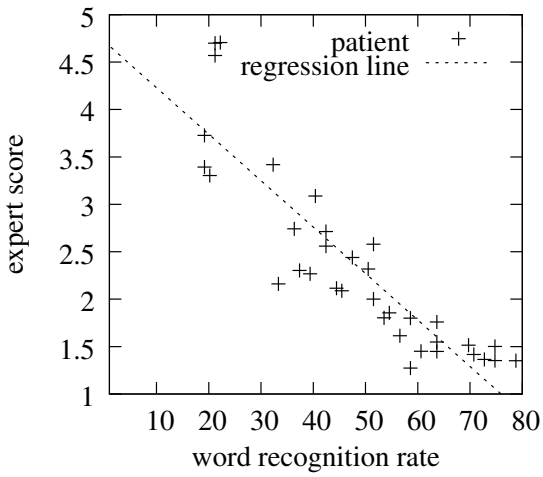

Figure 1: Word recognition rates in comparison to the scores of the human experts for the patient group ($r = -0.89$)

We can exploit this information to approximate a reference word chain. This reference word chain contains just the words which are shown on the slide. Unfortunately this is not sufficient to calculate a good word accuracy since most of the children use carrier sentences like "This is a ..." which are regarded as wrongly inserted words even if the recognition would be perfect. In order to avoid this problem we applied the word recognition rate instead since it does not weight the effect of inserted words. The difference between these methods is shown in Table 1.

6. Results

Since the control group was completely transliterated and recorded with our new software we could investigate the difference between the automatic measurements and those based on the transliteration. As can be seen in Table 2 the word recognition rate correlates to both transliteration-based measurements. The automatic word accuracy, however, matches poorly with the transliteration-based measurements (cf. Table 1). Therefore we expected the WR to show a good agreement with the results presented in (Maier et al., 2006).

The recordings of the CLP children showed a wide range of intelligibility (see Figure 1). Subjective speech evaluation showed good consistency. The correlation coefficient for the raters was 0.91. The results for the correlations of the WA, the WR and the subjective speech evaluation are shown in Table 3. When compared to the average of the raters, the WA for the recognizer has a correlation of -0.82 while the WR even correlates with -0.89. The coefficients are negative because high recognition rates come from "good" speech with a low score number and vice versa (note the regression line in Figure 1).

Figure 2 shows the word recognition rates of children in the same age range of both groups. As can be seen, almost all 45 children of the control group have high recognition rates. The distribution of the patients' group shows a high variance. This is due to the fact that the patients' group contained a wide range of intelligibility. Some of the patients were as intelligible as normal children (cf. Figure 1). The correlation between the age and the word recognition rate is 0.2 for the 45 children of the control group and 0.3 for the 20 children of the patient group. So there is just a weak connection between the age and the intelligibility.

7. Discussion

First results for an automatic global evaluation of speech disorders of different manifestations as found in CLP speech are shown. The speech recognition system shows high consistency with the experts' estimation of the intelligibility.The use of prior information about the speech test and its setup allows us the create a fully automated procedure to create a global assessment of the speaker's intelligibility. In difference to (Maier et al., 2006) no manual post-processing was done. Still the experts' and the recognizer's evaluation show a high correlation.

Using a control group we could show that our measure is sufficient to differentiate normal children's speech from pathologic speech. Furthermore we could show the consistency of our new measure to the transliteration-based evaluation methods.

The technique allows an objective evaluation of speech disorders and therapy effects. It avoids subjective influences from human raters with different experience and is therefore of high clinical and scientific value. Automatic

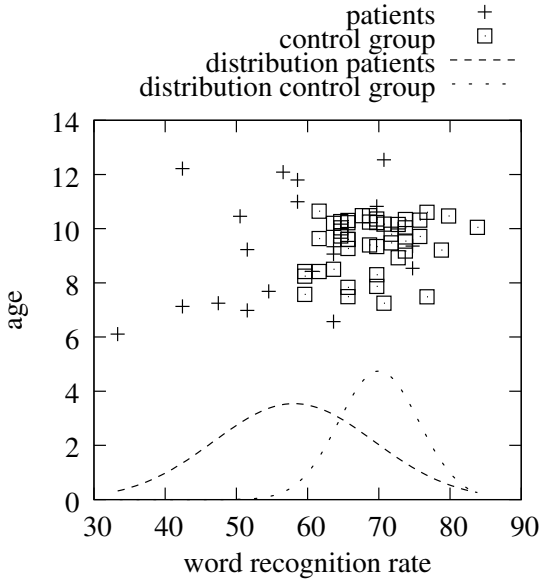

Figure 2: Distribution of the patients and the control group over the word recognition rate. Only members with about the same age were considered.

evaluation in real-time will avoid long evaluation proceedings by human experts. Further research will lead to the classification and quantification of different speech disorders. This will allow to quantify the impact of individual speech disorders on the intelligibility and will improve therapy strategies for speech disorders.

8. Conclusion

Automatic speech evaluation by a speech recognizer is a valuable means for research and clinical purpose in order to determine the global speech outcome of children with CLP. It enables to quantify the quality of speech. Adaptation of the technique presented here will lead to further applications to differentiate and quantify articulation disorders. Modern technical solutions might easily provide specialized centers and therapists with this new evaluation method.

9. Acknowledgments

This work was supported by the Johannes and Frieda Marohn foundation at the Friedrich-Alexander University of Erlangen-Nuremberg. Only the authors are responsible for the content of this article.

10. References

T. Bressmann, R. Sader, M. Merk, W. Ziegler, R. Busch, H.F. Zeilhofer, and H.H. Horch. 1998. Perzeptive und apparative Untersuchung der Stimmqualität bei Patienten mit Lippen-Kiefer-Gaumenspalten. *Laryngorhinootologie*, 77(12):700–708.

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

M. Gales, D. Pye, and P. Woodland. 1996. Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation. In *Proc. ICSLP '96*, volume 3, pages 1832–1835, Philadelphia, USA.

T. Haderlein, S. Steidl, E. Nöth, F. Rosanowski, and M. Schuster. 2004. Automatic recognition and evaluation of tracheoesophageal speech. In P. Sojka, I. Kopecek, and K. Pala, editors, *Text, Speech and Dialogue, 7th International Conference, September 8-11, 2004, Brno, Czech Republic, Proceedings*, volume 3206 of *Lecture Notes in Artificial Intelligence*, pages 331–338, Berlin, Heidelberg. Springer.

A. Harding and P. Grunwell. 1998. Active versus passive cleft-type speech characteristics. *Int J Lang Commun Disord*, 33(3):329–52.

T.T. Hogen Esch and P.H. Dejonckere. 2004. 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*, 68(8):1039–46.

C. Küttner, R. Schönweiler, B. Seeberger, R. Dempf, J. Lisson, and M. Ptok. 2003. Objektive Messung der Nasalanz in der deutschen Hochlautung. *HNO*, 51:151–156.

K. Van Lierde, M. De Bodt, J. Van Borsel, F. Wuyts, and P. Van Cauwenberge. 2002. Effect of cleft type on overall speech intelligibility and resonance. *Folia Phoniatr Logop*, 54(3):158–168.

A. Maier, C. Hacker, E. Nöth, E. Nkenke, T. Haderlein, F. Rosanowski, and M. Schuster. 2006. Intelligibility of children with cleft lip and palate: Evaluation by speech recognition techniques. In *Proc. International Conf. on Pattern Recognition*, volume 4, pages 274–277, Hong Kong, China.

T. Millard and L.C. Richman. 2001. Different cleft conditions, facial appearance, and speech: relationship to psychological variables. *Cleft Palate Craniofac J*, 38:68–75.

S. Paal, U. Reulbach, K. Strobel-Schwarthoff, E. Nkenke, and M. Schuster. 2005. Beurteilung von Sprechauffälligkeiten bei Kindern mit Lippen-Kiefer-Gaumen-Spaltbildungen. *J Orofac Orthop*, 66(4):270–278.

K. Pearson. 1896. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society of London*, 187:253–318.

F. Rosanowski and U. Eysholdt. 2002. Phoniatric aspects in cleft lip patients. *Facial Plast Surg*, 18(3):197–203.

R. Schönweiler and B. Schönweiler. 1994. Hörvermögen und Sprachleistungen bei 417 Kindern mit Spaltfehlbildungen. *HNO*, 42(11):691–696.

R. Schönweiler, J.A. Lisson, B. Schönweiler, A. Eckardt, M. Ptok, J. Trankmann, and J.E. Hausamen. 1999. A retrospective study of hearing, speech and language function in children with clefts following palatoplasty and veloplasty procedures at 18-24 months of age. *Int J Pediatr Otorhinolaryngol*, 50(3):205–217.

E. G. Schukat–Talamazzini and H. Niemann. 1991. Das ISADORA-System – ein akustisch–phonetisches Netzwerk zur automatischen Spracherkennung. In B. Radig, editor, *Mustererkennung 1991*, volume 290 of *Informatik Fachberichte*, pages 251–258, Berlin. Springer–Verlag.

M. Schuster, E. Nöth, T. Haderlein, S. Steidl, A. Batliner,

and F. Rosanowski. 2005. Can you Understand him? Let's Look at his Word Accuracy — Automatic Evaluation of Tracheoesophageal Speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA. IEEE Computer Society Press.

M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth. 2006. Evaluation of Speech Intelligibility for Children with Cleft Lip and Palate by Automatic Speech Recognition. *Int J Pediatr Otorhinolaryngol*, 70:1741–1747.

G. Stemmer, C. Hacker, S. Steidl, and E. Nöth. 2003. Acoustic Normalization of Children's Speech. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 1313–1316, Geneva, Switzerland.

G. Stemmer. 2005. *Modeling Variability in Speech Recognition*. Logos Verlag, Berlin.

W. Wahlster. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, New York, Berlin.

N. Wantia and G. Rettinger. 2002. The current understanding of cleft lip malformations. *Facial Plast Surg*, 18(3):147–53.