Intelligibility of Children with Cleft Lip and Palate: Evaluation by Speech Recognition Techniques

Andreas Maier[1], Christian Hacker[1], Elmar Nöth[1]
Emeka Nkenke[2], Tino Haderlein[3], Frank Rosanowski[3], Maria Schuster[3]
[1]Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg
Martensstraße 3, 91058 Erlangen, Germany
[2]Mund-, Kiefer- und Gesichtschirurgische Klinik, Universität Erlangen-Nürnberg,
Glückstraße 11, 91054 Erlangen
[3]Abteilung für Phoniatrie und Pädaudiologie, Universität Erlangen-Nürnberg
Bohlenplatz 21, 91054 Erlangen, Germany
Andreas.Maier@informatik.uni-erlangen.de

published in: The 18th International Conference on Pattern Recognition, Proceedings, pp. 274-277

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 objective means to determine and quantify the intelligibility does not exist. An automatic speech recognition system was applied to 31 recordings of CLP children who spoke a German standard test for articulation disorders. The speech recognition system was trained with normal adult speakers' and children's speech. A subjective evaluation of the intelligibility was performed by a panel of 3 experts and confronted to the automatic speech evaluation. The automatic speech recognition yielded word accuracies between 1.2 % and 75.8 % (48.0 % ± 19.6 %) with sufficient discrimination. It complied with experts' rating of intelligibility. Thus we show that automatic speech recognition serves as a good means to objectify and quantify global speech outcome of children with CLP.

1. Introduction

Cleft lip and palate (CLP) is the most common malformation of the head. It can result in morphological and functional disorders [19], whereat one has to differentiate primary from secondary disorders [10, 12]. Primary disorders include e.g. swallowing, breathing and mimic disorders. Speech and voice disorders [14] as well as conductive hearing loss that affect speech development [13], are secondary disorders. Speech disorders can still be present after reconstructive surgical therapy. 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/) [6]. 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 [11]. Until now, objective means exist only for quantitative measurements of nasal emissions [8, 9, 7] and for the detection of secondary voice disorders [1]. But other specific or nonspecific 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 by

Stemmer [16] was used. A commercial version of the recognizer is used in high-end telephony-based conversational dialogue systems (www.sympalog.com). The recognizer can handle spontaneous speech with mid-sized vocabularies of up to 10,000 words. In a first acoustic analysis, the speech recognizer converts spoken speech into a sequence of feature vectors which consist of 12 Mel-Frequency Cepstrum Coefficients (MFCC). The first coefficient is replaced with 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 (50 ms in total). The recognition is performed with semi-continuous Hidden Markov Models (SCHMMs). The codebook contains 500 Gaussian densities which are shared by all HMM states. The elementary recognition units are polyphones – an extension of the well-known triphone approach [15]. The polyphones were constructed for each sequence of phones which appeared more than 50 times in the training set.

We used a unigram language model to weigh the outcome of each word model. It was trained with the transliteration of the spoken tests (see below). Thus, the frequency of occurrence for each word in the used text was known to the recognizer. 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. The test set perplexity of the language model is 94 which is high enough to represent low intelligibility as low word accuracies (WA) and high intelligibility as high word accuracies.

The speech recognition system had been trained with acoustic information from spontaneous dialogues of the VERBMOBIL project [18] and normal children's speech. The speech data of non-pathologic children voices (23 male and 30 female) 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 of the children with CLP. All speakers were asked to speak "standard" German. 90 % of the training population (85 male and 47 female) 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 [17].

MAP adaptation [5] with the patients' data lead to further improvement of the speech recognition system.

3. Patients

Acoustic files were recorded from 31 children with CLP at the age from 4 to 16 years (mean 10.1 ± 3.8 years). 2 of them had an isolated cleft lip, 5 an isolated cleft palate, 20 a unilateral cleft lip and palate and 4 a bilateral cleft lip and palate. The examination was included in the regular outpatient examination of all children with CLP. Informed consent had been obtained by all parents of the children prior to the examination. All children were native German speakers, some using a local dialect.

The children were asked to name pictures that were shown according to the PLAKSS test [4]. This German test consists of 99 words. It includes all possible phonemes of the German language in different positions (beginning, center and end of a word). Furthermore, the children were asked to sustain all vowels and nasals and repeat 6 sentences from the "Heidelberger Rhinophonie Inventar" [20]. These consist of 5 sentences without nasal consonants and one only with nasal consonants ("Nenne meine Mama Mimi"). In this paper only the data from the PLAKSS test was taken into account. The 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.

4. Subjective evaluation

A panel of 3 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 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, not Cohen's "basic" kappa but the weighted multi-rater kappa by Davies and Fleiss [2] was used. It allows to compare an arbitrary number of raters and weighs the difference between the values of intelligibility or WA, respectively. Several problems occur when comparing the ratings of the human experts and a speech recognition system. First of all, the human ratings were made on a Likert scale while the word accuracy is a continuous measure within a completely different range. A mapping of the word accuracy to the Likert scale had to be defined, since the kappa value can only be computed on discrete data. We rounded the experts' average intelligibility scores to the next integer and set thresholds defined as intervals on the WA scale on the recognizer's results manually, so that the difference between the experts' scores and the scores derived from the WA values was minimal. The correlation, however, can be calculated directly between the WAs and Likert scores. Therefore it is used as a second measurement to show the consistency of the results.

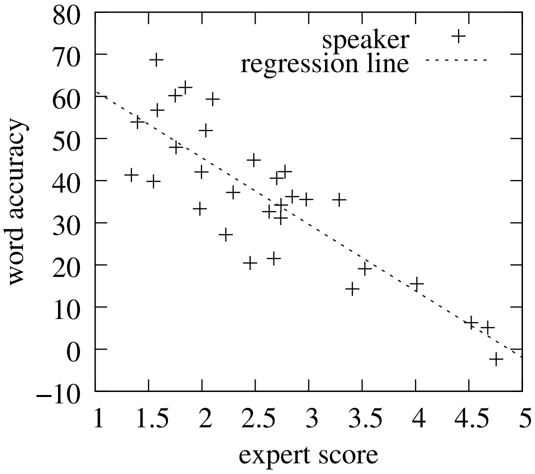

Figure 1. Word accuracies in comparison to the scores of the human experts (no adaptation)

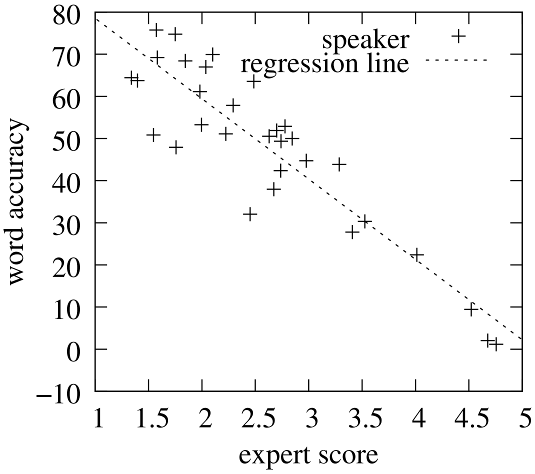

Figure 2. Word accuracies in comparison to the scores of the human experts (MAP adaptation)

Table 1. Inter-rater-correlation between the different raters

rater	K	S	L
other raters	+0.92	+0.93	+0.93

Table 2. Correlation between the different raters and the automatic speech recognizers (ASR)

rater	K	S	L	mean
ASR	-0.81	-0.85	-0.80	-0.85
adapted ASR	-0.84	-0.93	-0.85	-0.90

6. Results

The total duration of the children's audio files is 120 minutes, consisting of 5,330 words. The vocabulary of the word recognizer contains all 831 words occurring in the test data (99 unique words of the test, 34 words of the "Heidelberger Rhinophonie Inventar", 266 additional adjectives and alternative nouns which were used by the children to explain the pictures, and 432 additional words and word fragments representing reading errors). Due to the simple setup of the PLAKSS test the average turn length is very short (2.4 words). Because of the very limited amount of data we use the transliteration of all recordings as training set for the unigram language model. The vocabulary size is still large enough so that the acoustic realization of the children has high enough an influence on the word accuracy (the test set perplexity is 94). The recordings showed a wide range of intelligibility (see Figure 1). Subjective speech evaluation showed good consistency.The lowest correlation value between a rater and the mean of the other 2 raters is 0.92, the highest 0.93 (see Table 1). The results for the correlations of the WA and the subjective speech evaluation are shown in Table 2. When compared to the average of the raters, the WA for the recognizer has a correlation of -0.90 for the adapted case and -0.85 for the non-adapted case. 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 and Figure 2). The weighted multi-rater kappa for the group of the three raters is 0.48. The kappa for the rater group vs. the recognizer is 0.50 for the non-adapted case and 0.52 for the adapted case, i.e. the agreement between the human raters and the machine and the agreement among the humans alone can be regarded as identical (note that a result greater than 0.4 is said to represent fair agreement beyond chance [3]). Figure 2 shows the scores of human raters (averaged) and the adapted recognizer: the distance to the regression line is far better than in Figure 1 (cf. Table 2).

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 and sufficient discrimination of the intelligibility. Until now only few children with low speech intelligibility have been examined. Thus we have to collect more data and validate this technique with more patients. In this manner we can use the existing data as training set for the language model and evaluate with the new data. Thus we can create disjoint training and test sets.

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 evaluation in real-time will avoid long evaluation proceedings by human experts. Further research will lead to the recognition 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. The MAP adaptation seems to be beneficial for consensus between the recognizer and the experts.

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. It can easily be transposed into other languages. 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. We thank PD Dr. Peter Kummer and Dr. Jörg Lohscheller, Dpt. of Phoniatrics and Pedaudiology, University of Erlangen, for their fruitful comments on this work.

References

[1] 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*, 77(12):700–708, 1998.

[2] M. Davies and J. Fleiss. Measuring agreement for multinomial data. *Biometrics*, 38:1047–1051, 1982.

[3] J. Fleiss. *Statistical Methods for Rates and Proportions, 2nd ed.* John Wiley & Sons, New York, 1981.

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

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

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

[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*, 68(8):1039–46, 2004.

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

[9] 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 Phoniatr Logop*, 54(3):158–168, 2002.

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

[11] 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*, 66(4):270–8, 2005.

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

[13] R. Schönweiler, J. Lisson, B. Schönweiler, A. Eckardt, M. Ptok, J. Trankmann, and J. Hausamen. 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, 1999.

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

[15] E. G. Schukat–Talamazzini and H. Niemann. 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, 1991. Springer–Verlag.

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

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

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

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

[20] A. Zečević. *Ein sprachgestütztes Trainingssystem zur Evaluierung der Nasalität.* PhD thesis, Universität Mannheim, Germany, 2002.