INTERSPEECH 2009 BRIGHTON

Intelligibility Assessment in Children with Cleft Lip and Palate in Italian and German

Marcello Scipioni[1], *Matteo Gerosa*[2], *Diego Giuliani*[2], *Elmar Nöth*[3], *Andreas Maier*[3]

[1] Politecnico di Milano, Polo Regionale di Como, Italy
[2] FBK - Fondazione Bruno Kessler, Trento, Italy
[3] Chair of Pattern Recognition, Friedrich-Alexander-University Erlangen-Nuremberg, Germany
andreas.maier@informatik.uni-erlangen.de

Abstract

Current research has shown that the speech intelligibility in children with cleft lip and palate (CLP) can be estimated automatically using speech recognition methods. On German CLP data high and significant correlations between human ratings and the recognition accuracy of a speech recognition system were already reported. In this paper we investigate whether the approach is also suitable for other languages. Therefore, we compare the correlations obtained on German data with the correlations on Italian data. A high and significant correlation (r=0.76; $p < 0.01$) was identified on the Italian data. These results do not differ significantly from the results on German data ($p > 0.05$).

Index Terms: speech recognition, speech intelligibility, cleft lip and palate.

1. Introduction

Cleft lip and palate (CLP) is one of the most common alterations of the face. Clefts involving the lip and/or the palate are heterogeneous: they can involve only one structure, as cleft lip, or many of the structures in the face, like cleft lip and palate. Other kinds of cleft can also occur, like transversal and oblique clefts, which are called atypical clefts [1]. However, for simplicity, in this work we will refer to this group of anomalies as CLP.

The incidence of CLP is about of 1 per 1000, but can vary according to geographical location, ethnicity and socioeconomic status, with a higher prevalence among American Indians (3.6 per 1000) and Asians (1.5-2.0 per 1000) [2].

Speech of children with CLP shows particular characteristics, which can result in speech disorders also after surgical treatment. One of the most common disorders is hypernasality or nasal resonance, which is caused by resonance of sound in the nasal cavity. Hypernasality can be clearly perceived during production of vowels and diphthongs. Hypernasality can also be accompanied by nasal airflow, i.e. audible emission of air through the nose [3]. Another significant characteristic in speech of CLP children is the presence of errors in consonant production which comprise shifts in the point of articulation. The most typical consonant errors are pharyngeal backing, glottal articulation, absent or weakened pressure consonants and nasal fricatives [4]. These disorders affect the global intelligibility of CLP children.

In this context, the diagnosis of speech disorders is of crucial importance for the improvement of speech quality which would be a great help for speech therapists and important for documentation. In current clinical practice, speech evaluation is usually performed by speech therapists by means of auditive perception. This is very fast, but it is subjective and can lead to inaccurate evaluations, as shown in [5].

Our goal is to provide support to clinical procedures by means of an automatic assessment method, which is already in use in clinical practice in Germany. In this paper, we present the extension of this method from German to Italian language, to be able to generalize it to any language in the future.

2. Methods

2.1. State-of-the-art in Assessment of CLP Speech

The task of evaluation and assessment of speech data from a certain speaker—called test subject in the following—is to assign a label which corresponds to a certain property of the speaker's speech. Basically, this process is the same for any criterion but different scales and evaluation schemes can be chosen. In the literature many methods to evaluate disordered speech are found. In general, these can be divided into two groups:

- perceptual evaluations which are performed by a human subject—also called rater or labeler in this context—and
- objective evaluations obtained by a device or algorithm which is independent of a human rater.

"Objective measurement" of speech in a sense that it is also independent of the test subject is not possible for the case of speech evaluation because the test subject has to utter a sequence of words or vowels in all cases. Therefore, objective measurement of speech disorders in this context could also be called "instrument-based".

For the perceptual evaluation, many different methods and scales can be applied. The two main methods are quantification and qualification. For the quantification of different properties of speech, direct magnitude estimation and equal-appearing interval scales are widely used. The qualification of certain characteristics of speech is often done by classification of small parts of speech like phones or words. Therefore, often classes like "present" and "not-present" are chosen. However, experienced raters can even distinguish further grades in such small parts of speech as discussed in [6].

Objective means for CLP speech exist only for quantitative measurements of nasal emissions [7, 8, 9] and for the detection of secondary voice or speech disorders [10, 11]. But other specific or non-specific articulation disorders in CLP as well as a global assessment of speech quality cannot be sufficiently objectively quantified. In [12] an approach for the assessment of speech intelligibility in CLP children was presented for the German language. This approach used automatic speech recognition (ASR) to compute an estimate of the perceptual evaluation.

In this paper we investigate whether the approach of [13] is also suitable for the assessment of children with CLP in Italian language.

2.2. Italian ASR system setup

The ASR system for the Italian language was developed at FBK. Acoustic models were trained using the ChildIt speech corpus [14], augmented with additional read and spontaneous speech data. Speech data were collected from 179 normal children aged from 7 to 13 years (with an age average of 10) by using the same head-worn microphone and a sample frequency of 16 kHz. The overall duration of the training set, mostly formed by read speech, was about 11 hours.

Observation vectors, for Hidden Markov Model (HMM) based acoustic models, consisted of 39 acoustic features generated as follows. First, 13 Mel Frequency Cepstral Coefficients (MFCCs) were extracted using a 20ms Hamming window and an analysis step of 10ms. For each utterance, cepstral mean normalization was performed. Then, first and second order time derivatives of MFCCs were computed.

Both during training and testing, the 39 acoustic features were normalized on a speaker-by-speaker basis as follows. Acoustic observation vectors of a given speaker were normalized by applying an affine transformation whose parameters (i.e. transformation matrix and bias vector) were estimated through constrained maximum likelihood linear regression (MLLR; [15]) in order to maximize the likelihood of the speaker's data w.r.t. a *target* Gaussian mixture model (GMM). The GMM used as target model for transformation parameter estimation had 1024 components and was trained through a conventional ML training procedure by exploiting the unnormalized training data. With this normalization approach [16, 17], transformed/normalized feature vectors are supposed to contain less speaker, channel, and environment variability.

Speaker adaptive acoustic modeling was achieved by training recognition models on the normalized acoustic data through a conventional ML training procedure [16, 17]. Acoustic models for recognition were state-tied, cross-word triphone HMMs. In particular, a phonetic decision tree was used for tying the states of triphone HMMs. Output distributions associated with HMM states were modeled with mixtures with up to 32 diagonal covariance Gaussian densities, for a total of about 21000 Gaussian densities in the HMM set. Silence and several spontaneous phenomena were modeled with single-state HMMs.

For word recognition experiments we employed a word loop grammar containing all the words, 128, present in the target sentences. Similarly, for phoneme recognition experiments we employed a phoneme loop grammar containing 48 Italian phones.

2.3. German ASR system setup

For the objective measurement of the intelligibility of the German children with speech disorders, an automatic speech recognition system was applied, a 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 [18] was used. 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 MFCCs. 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 with full covariance matrices which are shared by all HMM states. The elementary recognition units are polyphones – an extension of the well-known triphone approach [19]. The polyphones were constructed for each sequence of phones which appeared more than 50 times in the training set.

A unigram language model was used to weigh the outcome of each word model. Thus, the frequency of occurrence for each word in the used text was known to the recognizer.

The speech recognition system had been trained with acoustic information from spontaneous dialogues of the VERBMOBIL project [20] 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 [21].

MLLR adaptation [15] with the patients' data lead to further improvement of the speech recognition system.

3. Patient Data

3.1. Italian Data

The Italian data set described here has been collected at the "Azienda Ospedaliera San Paolo", a hospital located in Milan. Some children in care at the Regional Center for Cleft Lip and Palate have been asked to take part at the data collection. Informed consent was obtained by the parents of the children. The patients have been recorded while uttering a set of sentences contained into a standardized test in use at the hospital.

The patients have been all recorded with a headphone set containing a microphone with an external Analog-to-Digital Converter (ADC), in order to get rid of possible distortions caused by the circuitry near the integrated sound card. The sampling frequency was 16 kHz.

The test in use at the hospital consists of 19 sentences, built in such a way to contain all the phonemes of the Italian language. Each sentence is focused on a specific consonant called target phoneme, which can appear at the beginning or in the middle of a word, combined with other phonemes or in groups of phonemes.

The patients are 12 children whose average age is 8 years. All the children are of Italian mother tongue, with the exception of one child; in this case his father's language is English. In most cases the pathologies of these children are cleft lip and/or palate: half of the patients have isolated cleft palate, while five present cleft lip and palate. Half of the recorded patients are male and half female. This means that a broad variety of cases are represented within the data set.

The collected data have been evaluated by an expert Italian speech therapist, who works by many years in the field of cleft lip and palate. The evaluation criterion is the global intelligibility of the patients. For each patient a global evaluation about intelligibility was given (0 = within normal limits, 1 = mild, 2 = moderate, 3 = severe) A lower value means thus

Table 1: Correlation between WA and the scores of the human experts. Both correlations do not differ significantly between languages ($p > 0.05$).

language	German	Italian
word accuracy	-0.86	-0.72

Table 2: Correlation between phoneme error rate and the scores of the human experts. Both correlations do not differ significantly between languages ($p > 0.05$).

language	German	Italian
phoneme error rate	0.86	0.76

better speech intelligibility, and a higher one indicates a higher severity of speech diseases.

3.2. German data

Acoustic files were recorded from 31 German children with CLP at the age from 4 to 16 years (mean 10.1 $\pm$ 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 out-patient examination of all children with CLP of the interdisciplinary cleft center located in Erlangen. 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 [22]. 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). 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.

A panel of five voice professionals perceptually 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.

4. Experiments and Results

The speech recognition was performed on word level and on phoneme level. The transliteration of the sentences uttered by the children were available. For the word level recognition the word accuracy (WA) was computed for the data in German and Italian language. The human expert score was available as mean of the evaluations performed by a panel of 5 experts for the German data, while for the Italian ones we have the score of a single human expert. The comparison between WA and the expert scores for the Italian data is shown in Figure 1. The correlations between WA and the expert scores are given in Table 1. For the phoneme level recognition, the phoneme error rate was computed for the target phonemes of the test. The correlations between the phoneme error rate and the expert scores are given in Table 2. Figure 2 shows the scatter plot for the Italian data ($r = 0.76; p < 0.01$).

5. Discussion and Future Work

The results obtained with the German data show a strong correlation with the average of the human scores. This strong correlation is confirmed also with the use of a panel of 5 raters, after that it was already shown in [23] with three raters. This method for intelligibility assessment is already in use in clinical practice in Germany for research purposes.

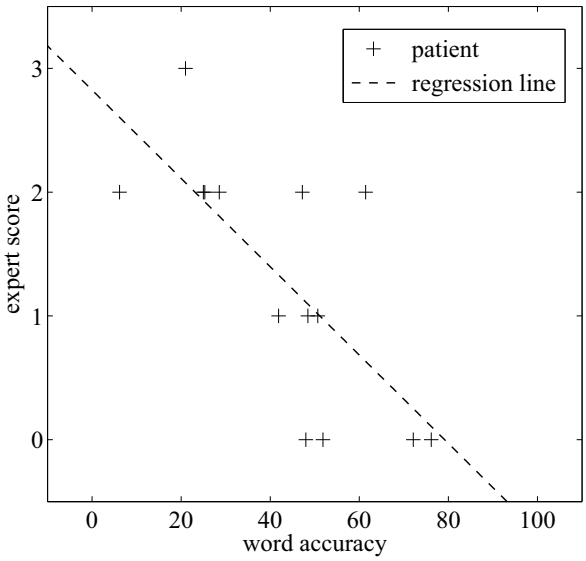

Figure 1: *Word accuracies in comparison to the scores of the human expert for the Italian data (r=-0.72; $p < 0.01$).*

The Italian dataset contains a smaller amount of patients than the German one, but a broad variety of cases is represented within the data set, as we can see in Figure 1. To compare the correlation between WA and the expert score obtained for the Italian data with the one obtained for the German data, we performed a significance test for correlation comparison (cf. [6, p.49]). The outcome was that the two correlations are not significantly different, with significance value $p > 0.05$ and considering the different number of children involved. Also comparing the correlation between the phoneme error rate and the expert scores for the Italian data with the one obtained for the German data, the significance test gave the same result.

Although the Italian data set is smaller than the German one and there was only one rater available for the Italian data, the results are consistent with the previous observations with the German data. Furthermore, the German data consisted of a single word naming task and the Italian data of read sentences. Still, there is no significant difference in terms of the correlations between the human experts and the speech recognition engine obtained so far. Therefore, we want to examine in further experiments if this still holds for more children and more raters.

As [24] also report high correlations for speech intelligibility with phonological features computed for the target phonemes, we would like to investigate this approach as well on our data. However, all these methods are language-dependent at least to some extent. Therefore, we need to determine the minimal set of data which is necessary to build an effective automatic speech intelligibility assessment tool, as we want to give instructions how to create such a tool for any language. To build such a system one will have to acquire normal speech data, pathologic speech data, and their annotations. In an optimal set, the required amount of annotated data should also be minimal because data annotation is the most expensive part of the data collection. For the normal data, read speech seems most promising as it can be transliterated fast since the reference text

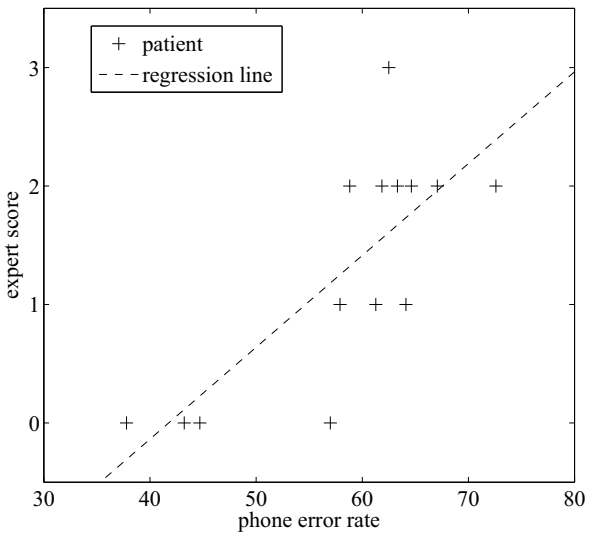

Figure 2: *Phoneme error rates in comparison to the scores of the human expert for the Italian data (r=-0.76; $p < 0.01$).*

is known. Furthermore, most speech tests are standardized and produce therefore either read or prompted speech. For the annotation procedure, interval scales also seem most beneficial as they are easy and fast to use. As reference they are in the same range in terms of correlation as other scales [25]. The final goal of our research is to provide automatic speech processing methods which are applicable for the speech intelligibility assessment in any language.

6. Acknowledgments

This project was performed with the collaboration of the Maxillofacial Division of the "San Paolo" Hospital of Milan and funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SCHU2320/1-1.

7. References

[1] M. T. Cobourne, "The complex genetics of cleft lip and palate," *European Journal of Orthodontics*, vol. 26, pp. 7–16, 2004.

[2] D. A. Nyberg, G. K. Sickler, F. N. Hegge, D. J. Kramer, and R. J. Kropp, "Fetal Cleft Lip with and without Cleft Palate: US Classification and Correlation with Outcome," *Radiology*, vol. 195, pp. 677–684, 1995.

[3] J. Bardach and H. L. Morris, *Multidisciplinary Management of Cleft Lip and Palate*. Saunders Company, 1990.

[4] 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.

[5] S. Paal, U. Reulbach, K. Strobel-Schwarthoff, E. Nkenke, and M. Schuster, "Evaluation of speech disorders in children with cleft lip and palate," *J Orofac Orthop*, vol. 66, no. 4, pp. 270–278, 2005.

[6] A. Maier, *Speech of Children with Cleft Lip and Palate: Automatic Assessment*. Berlin, Germany: Logos Verlag, 2009.

[7] 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.

[8] 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.

[9] 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. 1039–1046, 2004.

[10] 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.

[11] A. Zečević, "Ein sprachgestütztes Trainingssystem zur Evaluierung der Nasalität," Ph.D. dissertation, University Mannheim, Germany, 2002.

[12] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS – A System for the Automatic Evaluation of Voice and Speech Disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[13] A. Maier, E. Nöth, A. Batliner, E. Nkenke, and M. Schuster, "Fully Automatic Assessment of Speech of Children with Cleft Lip and Palate," *Informatica*, vol. 30, no. 4, pp. 477–482, 2006.

[14] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, pp. 847–869, 2007.

[15] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[16] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization." *Computer Speech and Language*, vol. 20, no. 1, pp. 107–123, Jan. 2006.

[17] G. Stemmer, F. Brugnara, and D. Giuliani, "Using Simple Target Models for Adaptive Training," vol. 1, Philadelphia, PA, March 2005, pp. 997–1000.

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

[19] E. G. Schukat–Talamazzini and H. Niemann, "Das ISADORA-System – ein akustisch–phonetisches Netzwerk zur automatischen Spracherkennung," in *Mustererkennung 1991*, ser. Informatik Fachberichte, B. Radig, Ed., vol. 290. Berlin: Springer–Verlag, 1991, pp. 251–258.

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

[21] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, "Acoustic Normalization of Children's Speech," in *Proc. European Conf. on Speech Communication and Technology*, vol. 2, Geneva, Switzerland, 2003, pp. 1313–1316.

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

[23] 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.

[24] C. Middag, G. Van Nuffelen, J. Martens, and M. De Bodt, "Objective intelligibility assessment of pathological speakers," in *Interspeech 2008 – Proc. Int. Conf. on Spoken Language Processing, 11th International Conference on Spoken Language Processing, September 25-28, 2008, Brisbane, Australia, Proceedings*, 2008, pp. 1175–1178.

[25] T. Haderlein, K. Riedhammer, A. Maier, E. Nöth, H. Toy, and F. Rosanowski, "An Automatic Version of the Post-Laryngectomy Telephone Test," in *10th International Conf. on Text, Speech and Dialogue (TSD)*, ser. Lecture Notes in Artificial Intelligence, V. Matoušek and P. Mautner, Eds., vol. 4629. Berlin, Heidelberg, New York: Springer, 2007, pp. 238–245.