Towards a Language-independent Intelligibility Assessment of Children with Cleft Lip and Palate

Tobias Bocklet
Chair of Pattern Recognition
Martensstr. 3
91058 Erlangen, Germany
tobias.bocklet@informatik.uni-erlangen.de

Andreas "Magnet" Maier
Department of Pedaudiology and Phoniatrics
Bohlenplatz 21
91054 Erlangen, Germany
andreas.maier@informatik.uni-erlangen.de

Korbinian Riedhammer
Martensstr. 3
91058 Erlangen, Germany
sikoried@i5.informatik.uni-erlangen.de

Elmar Nöth
Chair of Pattern Recognition
Martensstr. 3
91058 Erlangen, Germany
noeth@informatik.uni-erlangen.de

ABSTRACT

We describe a novel evaluation system for the intelligibility assessment of children with CLP on standardized tests. The system is solely based on standard cepstral features in form of MFCCs. No other information like word alignments is used. So the system can be easily adapted to other languages. For each child one GMM is created by adaptation of a UBM to the speaker-specific MFCCs. The components of this GMM are concatenated in order to create a so-called GMM supervector. These GMM supervectors are then used as meta features for an SVR. We evaluated our language-independent system on two different datasets of children suffering from CLP. One dataset contains recordings of 35 German children, where the children named different pictograms. The other dataset contains recordings of 14 Italian speaking children, who repeated standardized sentences. On both datasets we achieved high correlations: up to 0.81 for the German dataset and 0.83 for the Italian dataset.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Process.

General Terms

Languages,Human Factors

Keywords

Gaussian Mixture Models, Support Vector Regression, Acoustic Analysis, Cleft Lip and Palate

ICMI-MLMI'09 Workshop on Child, Computer and Interaction November 5, 2009, Cambridge, MA, USA

1. INTRODUCTION

Cleft lip and palate (CLP) is the most common malformation of the head with incomplete closure of the cranial vocal tract. Speech disorders can still be present after reconstructive surgical treatment. The characteristics of speech disorders are mainly a combination of different articulatory aspects, e.g., 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 the back of the tongue or vice versa, and a modified articulatory tension, e.g., weakening of the plosives. They affect not only the intelligibility but therewith the social competence and emotional development of a child.

For speech therapy of children with CLP, it is very important to evaluate the development of a child's voice before and after reconstructive surgery. This is normally achieved by an intelligibility assessment of a speech therapist familiar with children's speech. Objective, automatic measures are very helpful for this task. In [8] a system based on automatic speech recognition (ASR) has been shown to be useful for this task: We achieved high correlations (0.9) between the word accuracy (WA) of the system and the intelligibility scores of human experts. A drawback of this system is it's complexity. If one wants to adapt it to another language a lot of training speech data and the corresponding transliterations are needed.

In this work we present an approach which is solely based on acoustic features. The idea behind the approach rests upon the assumption that there is a similarity between automatically computed acoustic properties and the intelligibility scores from expert listeners. The acoustic space of a child is modeled by Gaussian Mixture Models (GMMs). Therefore, a Universal Background Model (UBM) is adapted by Maximum A Posteriori (MAP) adaptation to speaker-specific spectral features. This creates a speaker-specific GMM, that models the acoustic space of a certain child. In order to evaluate the acoustics of a child correctly/adequately, it is important that every child speaks the same text or a text, that contains all phonemes of a certain language. The idea is now, that children with the same severeness of CLP and in-

telligibility have similar acoustic spaces, i.e., similar GMMs, so that GMMs can directly be used to assess the intelligibility. Based on these GMMs, supervectors are extracted. Supervectors consist of certain components of the GMMs. These supervectors are labeled with the experts' intelligibility score and are used as input vectors for a Support Vector Regression (SVR). The remainder of the paper is organized as follows. Section 2 introduces the datasets used for training and evaluation. In order to show, that the system can also be used for different languages, we evaluated it on a dataset of Italian and a dataset of German children. Section 3 describes the evaluation system and the different employed GMM-based supervectors. In Section 4 we present leave-one-speaker-out results for the two different datasets. The paper concludes with a summary (Section 5) and proposed future research (Section 6).

2. DATASETS

Two different datasets have been used in this paper. Each dataset was recorded with a sampling rate of 16 kHz and a quantization of 16 bit. The recordings were performed with PEAKS (*Program for the Evaluation and Assessment of all Kinds of Speech*) [7].

2.1 German CLP dataset

35 German children with CLP were recorded at the Department of Pedaudiology and Phoniatrics of the University Hospital in Erlangen. The mean age of the children was 8.5 ± 3.5 years. All children performed the PLAKSS Test [4], a semi-standardized test which is commonly used by speech therapists. The test is composed of 99 pictograms. These have to be named by the children. The test contains all phonemes of the German language and the most important conjunctions among them at different word positions (beginning, central or ending). All children were recorded with the same microphone, a standard headset microphone (dnt Call 4U Comfort) with external Analog-to-Digital-Converter (ADC). The mean recording time was approximately 9 minutes.
For each child an 5-scale Likert-based intelligibility assessment as described in [6] was used. The subjective intelligibility assessment was performed by 5 speech therapists, who are familiar with CLP children.

2.2 Italian CLP dataset

14 Italian children with CLP were recorded at the Maxillofacial Division of the "Azienda Ospedaliera San Paolo", a hospital located in Milan. All children are native Italian speakers. The mean age of the 14 children was 8 years. All children repeated a set of sentences that are part of a standardized test. The test was developed at the San Paolo hospital and consists of 19 sentences. Each sentence focuses on a certain phoneme. The phoneme can appear at the beginning or in the middle of a word. Combinations with other phonemes or groups of phonemes are also handled in these sentences. The mean recording time is approximately 3.5 minutes.
All children were recorded with the same microphone with an external ADC. A global intelligibility evaluation by an expert Italian speech therapist was performed on the dataset. This evaluation is similar to the evaluation of the German dataset, but the scale ranged from 0 to 3. A lower value means thus better speech intelligibility. A higher one indicates a lower speech intelligibility.

3. AUTOMATIC ASSESSMENT SYSTEM

In this paper we employ an evaluation system which is based on Support Vector Regression (SVR) and Gaussian Mixture Models (GMM). We first published the system in [2] where it was used to determine the age of children. For that kind of evaluation, we assumed that the acoustics are changing during the growth of the vocal tract. We based our evaluation only on a shift of the mean vectors from the UBM to the mean vectors of a speaker GMM. In [1] we adapted the system to the task of intelligibility assessment of patient's with partial laryngectomy. Our main assumption was that the acoustics of a pathologic speaker differ to the acoustics of a normal speaker. Normal speakers have the possibility to pronounce a vowel or word always in the same way. Pathologic speakers have difficulties in doing so. Thus, we used not only the mean vectors but also the variances to perform an evaluation of pathologic speech. We achieved correlations to human raters of up to 0.83 with this system.

In this section we briefly describe the system. First, short-time features in form of Mel-Frequency Cepstrum Coefficients (MFCCs) are extracted from the speech signal. GMMs are employed to model these features for each child, i.e., a GMM for each child is adapted from a UBM. This GMM models the acoustic space of a child. In our system the GMM is then used as a meta feature vector for an SVR. Therefore different components of a GMM are concatenated to a huge vector. This vector is called GMM-based supervector. We evaluated different kinds of supervectors.

3.1 Feature Extraction

As features we use the well-known MFCCs. These features perform a short time analysis of the speech signal. Therefore a Hamming window with a size of 16ms and a time shift of 10ms is applied to the signal. The Mel spectrum with triangle filters is calculated afterwards. The cepstral coefficients are computed by an inverse discrete cosine transform of the logarithmic Mel spectrum. Finally, a 24-dimensional feature vector is created. It contains the first 12 Mel-frequency cepstral coefficients and their first-order derivatives.

3.2 GMM-based Supervectors

GMMs model the acoustic features, and with this the acoustic space, of a specific speaker. A GMM (λ) contains M unimodal Gaussian densities. Each density represents a different acoustic area:

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

$$= \sum_{i=1}^{M} \omega_i \cdot \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma_i}|^{(1/2)}} e^{-(1/2)(\boldsymbol{c}-\boldsymbol{\mu_i})^T \boldsymbol{\Sigma_i}^{-1}(\boldsymbol{c}-\boldsymbol{\mu_i})}, \quad (2)$$

The idea is now to train these speaker-specific GMMS and use their parameters in a Support Vector Regression (SVR) [9]. The parameters of the densities i ($i = 1, ..., M$), i.e., weight (w_i), mean ($\boldsymbol{\mu_i}$) and covariance ($\boldsymbol{\Sigma_i}$) are concatenated to a vector afterwards.

3.2.1 GMM Training

After feature extraction a UBM is created on a dataset of healthy speakers. This is achieved by using 5 iterations of

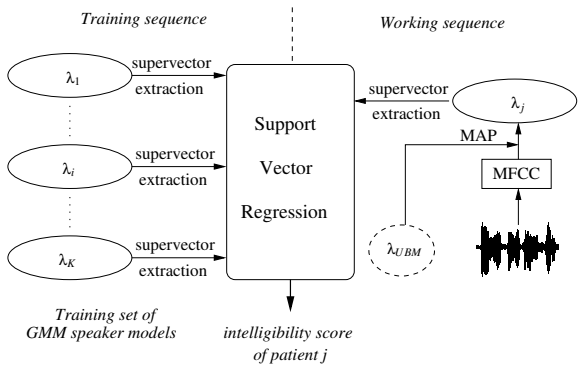

Figure 1: Principle of the Support Vector Regression (SVR) system

the EM algorithm [3]. Beginning with this UBM a speaker-dependent GMM is built by MAP adaptation [5]. The MAP adaptation takes the UBM as an initial model and adapts the statistics to the acoustic features of a specific speaker in a single iteration step. These new densities are combined with the UBM statistics afterwards. The combination of new and old statistics is based on a relevance parameter, which takes the number of speaker-specific feature vectors into account. Finally, a GMM is created for each speaker. We did not observe major differences between full and diagonal covariance matrices. Because of that we only present the results achieved with diagonal covariance matrices, and the notation refers to the values of a diagonal covariance matrix.

3.2.2 *Final Support Vector Regression System*

Figure 1 shows the principle of the GMM supervector-based SVR system. For each child of the training set, a GMM λ is created. The components of each GMM are concatenated to a GMM-based supervector. These supervectors are used as input vectors for the SVR training. They can be regarded as a mapping from the acoustics of a speaker, i.e., MFCCs, to a higher-dimensional feature vector which represents the speaker himself or his/her characteristics. In the case of this paper, the property of interest is the intelligibility. If the intelligibility of a speaker j has to be determined, a speaker model λ_j is derived from the background model λ_{UBM}. Note that the background model is the same that is used to create the GMMs of the training speakers. A GMM supervector is created for speaker j, and an intelligibility prediction is performed by SVR. For our experiments we employed an SVR with a linear kernel.

4. EXPERIMENTS AND RESULTS

We evaluated our system on two different datasets of children with CLP. A database of 35 German speaking children, and a database of 14 Italian speaking children. Due to lack of training data, we used an UBM trained on healthy German children with a similar age distribution. We based the experiments on different supervectors. In [1] we varied the composition of the GMM supervectors. Four of them have been of bigger interest. The experiments in this work are based on the 4 different types of supervectors: one containing the weights, one containing the mean vectors, one containing the diagonal covariance matrices and finally a combination of all three of them. On the German dataset we achieved correlation to the mean value of the 5 human experts of up to $r = 0.81$. On the Italian dataset a maximum correlation coefficient of $r = 0.83$ was achieved.

# of densities	Context 3 18 Mel filters	Context 3 25 Mel filters	Context 5 25 Mel filters
German			
32	0.75	**0.79**	0.71
64	0.72	**0.80**	0.70
128	0.75	**0.80**	0.73
Italian			
32	0.76	0.64	**0.76**
64	0.68	0.61	**0.71**
128	0.53	0.55	**0.61**

Table 1: Pearson correlations against different MFCC parameters on the two datasets

4.1 Results with different MFCC parameters

We investigated different parameters of the feature extraction process. We varied the number of filters in the Mel spectrum and the context length for the derivative calculation. The experiments have been performed for both datasets. We did not find significant difference for the different parameters within one dataset. But we found regularity within these parameters for each of the two datasets: On the German dataset the best results where achieved, when the first derivative is calculated with a context of three and a Mel spectrum with 25 triangular filters. On the Italian dataset a context of five and a Mel spectrum with 25 triangular filters achieved the best results. These regularities could arise because of the different languages and the different types of recordings. The German dataset contains recordings, where children have to name small pictograms. The Italian dataset contains repeated sentences. A more general statement needs a deeper investigation with bigger datasets. In Table 1 a summary for the MFCC parameters is given. These experiments used a supervector containing weight, means and diagonal covariance matrices.

4.2 Results on German dataset

The results achieved on the German dataset are shown in Table 2. The experiments are based on different supervectors. The first column contains supervectors, where only the weights are incorporated. Here, the best correlation coefficient of 0.62 was achieved with 256 Gaussians. Supervectors that contain the mean vectors achieved the best result of 0.79 also with 256 Gaussian densities. A supervector composed of the diagonal covariance entries could not reach this value. With a growing number of densities the correlation coefficient is growing, especially when focusing on supervectors that contain weights or mean vectors. More densities lead to a more precise observation of the acoustic properties of the children's speech. Covariance supervectors reach a stable result of 0.72 or 0.73. Due to the high amount of available data for each speaker (approx. 9 min.) the dataset is quite balanced and GMMs with a high number of Gaussian mixtures can be trained.
The combination of all three Gaussian components to a very high-dimensional vector a correlation coefficient of 0.81 could be achieved. 256 Gaussians were used in this case.

# of Gaussians	supervectors			
	w	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	$w+\boldsymbol{\mu}+\boldsymbol{\Sigma}$
32	0.62	0.67	0.72	0.79
64	0.54	0.75	0.72	0.80
128	0.51	0.77	0.73	0.80
256	0.71	0.79	0.73	**0.81**

Table 2: German dataset: Pearson correlation dependent on the number of Gaussian densities and the components of the supervector

# of Gaussians	supervectors			
	w	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	$w+\boldsymbol{\mu}+\boldsymbol{\Sigma}$
32	0.42	0.69	**0.83**	0.76
64	0.35	0.63	0.78	0.71
128	0.42	0.53	0.51	0.61
256	0.36	0.49	0.47	0.42

Table 3: German dataset: Pearson correlation dependent on the number of Gaussian densities and the components of the supervector

The dimension of the supervector is 12544 in this case. The differences between the four different kinds of supervectors are not significant at any comparison.

4.3 Results on Italian dataset

The correlation coefficients achieved on the Italian dataset are summarized in Table 3. Supervectors containing the weights achieved only low correlations of up to 0.42 with 32 or 128 Gaussian densities. A significant improvement ($p \leq 0.1$) compared to weight-based supervectors could be achieved with 768 dimensional (32 Gaussians) mean supervectors. The correlation coefficient is 0.69 in this case. Supervectors using diagonal covariance matrices achieved $r = 0.83$. Due to the low number of speakers, the improvement is not significant compared to the previous results. A combination of all three components achieved an inferior correlation coefficient of 0.76.
It can be observed that a higher amount of Gaussian densities did not lead to higher correlation coefficients. Because of the low amount of speech data for each speaker (approx. 3.5 min) no better modeling of the acoustic space was achieved when increasing the number of densities. This effect is amplified by the low amount of test speakers in the Italian dataset, i.e., 14 speakers, so that a balanced training of the SVR could not be ensured.

5. SUMMARY

In this work we presented a system for the analysis and assessment of voices of children suffering from CLP. The system is only based on acoustic information, so that it is mostly likely working for different languages. In order to show the language independent character of the system, we evaluated it on two different datasets: 35 German children and 14 Italian children. The system models the acoustic space of the children by GMMs. The main assumption is, that there is a linear relation between these GMMs and the intelligibility score assigned by human experts. On the German dataset we achieved a correlation between these experts' score and the GMM of $r = 0.81$. On the Italian dataset our highest achieved correlation was $r = 0.83$.

6. FUTURE WORK

We have shown, that the introduced system can be easily adapted to different languages. This is a first step towards a language independent evaluation and assessment system of CLP children. In future work we would like to focus on datasets with comparable expert's scores.
Additionally we would like to focus on a more detailed analysis of the proposed feature extraction mechanism and the corresponding parameters.

7. ACKNOWLEDGMENTS

The authors thank Augusto Sarti and Marcello Scipioni of the *Dipartimento di Elettronica e Informazione - Politecnico di Milano* for the provision of the Italian dataset and the corresponding intelligibility labels.

8. REFERENCES

[1] T. Bocklet, T. Haderlein, F. Hönig, and E. Nöth. Evaluation and Assessment of Speech Intelligibilty on Pathologic Voices Based upon Accoustic Speaker Models. In *Proceedings of the 3rd Advanced Voice Function Assessment International Workshop*, pages 89–92, Madrid, 2009.

[2] T. Bocklet, A. Maier, and E. Nöth. Age Determination of Children in Preschool and Primary School Age with GMM-Based Supervectors and Support Vector Machines/Regression. In P. Sojka, A. Horak, I. Kopecek, and K. Pala, editors, *Proceedings Text, Speech and Dialogue; 11th International Conference*, Lecture Notes in Artificial Intelligence, pages 253–260, Heidelberg, 2008.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

[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] T. Haderlein, E. Nöth, H. Toy, A. Batliner, M. Schuster, U. Eysholdt, J. Hornegger, and F. Rosanowski. Automatic Evaluation of Prosodic Features of Tracheoesophageal Substitute Voice. *Eur Arch Otorhinolaryngol*, 264(11):1315–1321, 2007.

[7] 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*, 51(5):425–437, 2009.

[8] 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*, 70:1741–1747, 2006.

[9] A. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.