Speech Classification for Sigmatism in Children

Introduction

Sigmatism, also known as lisping, is a speech disorder associated with the misarticulation of sibilants phonemes. In the German language it is observed in the alveolar sibilants /s/ and /z/. If the tongue or lips are not in the correct position or shape then the air flow changes and the person produces sounds different to those intended. The intelligibility of the speech is not highly damaged [1], but speech perception is altered [2].

Depending on the tongue position, we can distinguish different types of sigmatism, such as interdental and lateral. Interdental lisps happens when the tongue stretches out between the front teeth creating a // sound instead of a /s/ or /z/ and lateral occur when the tongue touches the roof of the mouth and the air is pushed outwards laterally.

Sigmatism can appear in previously unaffected children during the second dentition. It is usually only considered a disorder if it occurs in children older than six. Treatment is assisted through speech therapy sessions where the patient is taught how to position and stress both tongue and lips for the correct pronunciation of the target phonemes. However, it is problematic for the patient to practice it alone since their own perception is also impaired. Therefore it is important to assist treatment with a tool that provides the patient with feedback even without the presence of a speech therapist.

Previous attempts were made to assist lisping speech therapy using visual aid as well [3], but only a few studies were performed with the pathological speech data that could have lead to a complete automatic evaluation system for this disorder. Other child speech disorders, such as hypernasality and pharyngealization, were already automatically evaluated through the use of word accuracy and spectral cues [4]. The acoustic study on lateral sigmatism in Japanese [5] shows that spectral envelope peaks are found in areas different to those for normal speech. A classification system for sigmatism in Arabic described in [6] however uses Mel Frequency Cepstrum Coefficients (MFCCs) for this classification task.

In this paper we first present results of an acoustic study performed over the available pathological data. Further on we discuss features sets using MFCCs and energy, and present the evaluation results of each classification scheme.

Speech Data and Acoustic Study

A group of 26 children with cleft lip and palate (5 female and 21 male) were recorded using a standard head set (dnt Call 4U Comfort) speaking the PLAKSS Test [7], a German semi-standardized test commonly used in speech therapy. The data was sampled at 16 kHz with a 16 bit quantization. The dataset used for this work is a subset of the data which has already been investigated in [8] for other speech disorders.

Labels	Phones	Children
normal	431	25
sigmatism interdental	129	7
sigmatism lateral	80	12

Table 1: The labeling given by the speech therapist to each phone appearing in the dataset of 26 children, and the number of children appearing in each class.

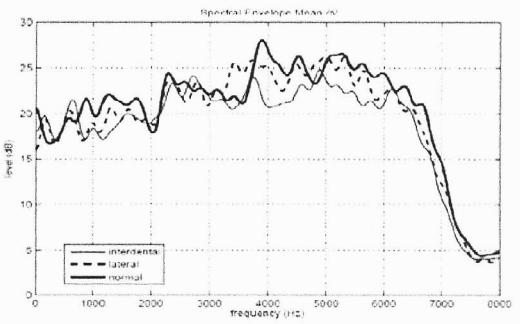

Fig. 1: Spectral envelope of /s/ for the same speaker.

Our dataset is composed of 640 phones that are realizations of /s/ and /z/ in different contexts, such as /ts/ ('zitrone', in English lemon) and /ks/ ('hexe', in English witch). The data was manually transliterated and was segmented through the use of an automatic speech recognition system.

Table 1 displays the different labels given by the speech therapist to each phone, as well as number of phones in each class. The total number of children in each class does not add to the overall number of children in the dataset because some children had normal and pathological realizations of some phones, appearing therefore in more than one class.

For the acoustic study of the pathology we processed the speech signal frame by frame using a hamming window of 25.6ms length at a rate of 100Hz. Each frame was filtered with an FIR high pass filter. Afterwards, we estimated the spectral envelope through cepstrum smoothing, using the true estimator

technique [9]. All phones were normalized by the maximum absolute amplitude within the phone since we believe that it is not a discriminative feature and that normalization would decrease the deviation of the spectral mean results.

The envelope was evaluated for each of the three labeled data set, so as to lead us to the relevant spectral clues for the classification task. Figure 1 shows the spectral envelope of a certain speaker that was labeled in all three cases for the pronunciation of /s/. Figure 2 shows the mean spectral envelope for each label for all speakers and phones.

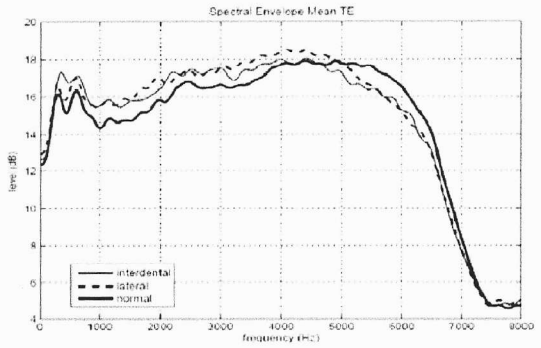

Fig. 2: Mean of spectral envelope of all phones for all speakers.

One can see that in both figures a strong damping appears in all curves for the higher frequencies due to the anti-aliasing low pass filter operation carried out during the recording of our dataset. As the sampling frequency was 16 kHz the digital representation of the signal could only contain spectrum in areas up until 8 kHz.

As expected, Figure 1 shows that the normal speech envelope has on average higher levels than either of the pathological envelopes, more clearly for frequencies above 5 kHz.

The deviation in the mean curves displayed in Figure 2 was around 4 dB for each sample. Even with the normalization that we applied, a high deviation was found. This occurred because of segmentation problems in the speech database and due to inter speaker and inter phone differences as well. But even considering the appropriate deviation intervals it is still possible to distinguish that, as happened with the /s/ sound from the previous example, higher levels are present in normal speech in the higher frequency bandwidths.

Classification Schemes

Considering the results we retrieved with the acoustic analysis, we choose three different feature sets:

1. MFCCs: widely used to characterize manners of articulation for speech recognition and other classification tasks for speech processing. In our task we calculated 12 coefficients for each frame and for each phone the mean of the coefficients was retrieved as the feature vector of that phone;

2. ModifiedMFCCs: instead of filtering the cepstrum representation of speech using mel filter banks, we used mirrored versions of the whole filter bank structure in order to increase frequency resolution in high frequency areas, that are the most discriminative for our task. Once again, this feature vector contains the mean value of the 12 coefficients of each frame;

3. Energy: total energy within two frequency bandwidths of the speech signal. According to what is shown in Figure 2, we choose the following bandwidths: 1-4 kHz and 5-6 kHz. The feature vector is composed of two energies calculated over the mean of the spectral envelope of the phone.

We selected the following classifiers from the various provided by the WEKA toolbox [10]:

4. NaivesBayes: the class conditionals are modeled as a unimodal Gaussian distribution;

5. Logistic: the class conditionals are modeled as a unimodal Gaussian distribution, with same covariance matrix, and the classification decision function is the logarithm of the ratio of the posterior;

6. ViaRegression: one linear regression model is built for each class value.

Experimental Results

It is not our goal to distinguish which type of sigmatism is present, but to indicate whether a sigmatism is present at all. Therefore both the interdental and lateral dataset were considered as belonging to the same class. For training the classifiers a 10 fold cross validation procedure was applied.

The measures we used to compare the performance of each scheme are the absolute recognition rate (RR), the class-wise averaged recognition rate (CL) and the area under the receiver operating characteristics curve (AUC) [11]. The results are displayed for each feature set in the Tables 2, 3 and 4.

The RR indicator is computed as the number of correctly recognized phones (from both classes) divided by the number of all tested phones. And the CL is determined as the average of the recognition rate per class, also referred to as "recall". It is important to have both RR and CL results since the distribution of the classes in the data set is not balanced, which biases the

RR calculation. However the CL alone is also not representative.

Classifier	RR (%)	CL (%)	AUC
NaivesBayes	63.4	**54.5**	0.63
Logistic	65.8	**54.7**	**0.64**
ViaRegression	**66.1**	54.0	0.62

Table 2: MFCCs feature set

Classifier	RR (%)	CL (%)	AUC
NaivesBayes	64.3	**55.5**	**0.65**
Logistic	**66.1**	54.9	0.64
ViaRegression	**66.1**	53.7	0.61

Table 3: Modified MFCCs feature set

Classifier	RR (%)	CL (%)	AUC
NaivesBayes	66.9	52.3	0.63
Logistic	**67.3**	**54.6**	0.65
ViaRegression	67.2	54.2	**0.66**

Table 4: Energy feature set

The AUC is the area under the ROC curve, and tells us the overall behavior of the classifier regardless of the cost value for the misclassification of each class. When we consider a two class classifier, one can rank the instance probabilities, i.e. the degree to which an instance is a member of a class. The AUC measure is equivalent to the probability that the classifier will position in this rank a randomly chosen positive instance higher than a randomly chosen negative one [11].

The results indicate that, although the performance of classifiers using the same feature set do not show significant differences, the overall result of each feature set is more noticeable and the energy feature set, of only two dimensions, showed better classification results with respect to the AUC criteria than the MFCC based sets.

Conclusions

This study presents preliminary results obtained using a German language speech database composed of children with a variety of pathological disorders for the purpose of sigmatism classification. A spectral acoustic study was performed and showed us that the energy level for the higher frequency bands is different for the normal and the pathological cases. This finding guided us to experiment with direct calculation of spectral energy as a feature set. From the evaluation results of the classifiers we could observe slightly better performance of the energy feature set as compared to the baseline approach of using MFCCs as feature vectors. Further works will investigate a more detailed model for the pathology by analyzing simulated pathological speech data created by specialists. This dataset will be recorded at a higher sampling frequency (44.1 kHz). It will not only give us the information of higher frequency bands that were either damped or not present at all in our dataset, but also provide us with a model of typical sigmatism in the opinion of the speech therapist.

Literature

[1] A. Maier, "Speech of Children with Cleft Lip and Palate: Automatic Assessment," PhD Thesis Pattern Recognition Chair Friedrich-Alexander Universität Erlangen Nürnberg, pp. 109–110, 2009.

[2] D. E. Mowrer, P. Wahl and S. J. Doolan, "Effect of lisping on audience evaluation of male speakers," Journal of Speech and Hearing Disorders, vol. 43, pp. 140-148, 1978.

[3] K. Grauwinkel and S. Fagel, "Visualization of internal articulator dynamics for use in speech therapy for children with Sigmatismus Interdentalis," International Conference on Auditory-Visual Speech Processing, paper P32, 2007.

[4] A. Maier, F. Hönig, C. Hacker, M. Schuster and E. Nöth, "Automatic Evaluation of Characteristic Speech Disorders in Children with Cleft Lip and Palate," Interspeech 2008, vol. 1, pp. 1757–1760, 2008.

[5] Akagi M, Suzuki N, Hayashi K and Saito H, "Perception of Lateral Misarticulation and Its Physical Correlates," Folia Phoniatr Logop, no. 6, vol. 53, pp. 291–307, 2001.

[6] Z. A Benselama, M. Guerti and M.A. Bencherif, "Arabic Speech Pathology Therapy Computer Aided System," Journal of Computer Science, no. 9, vol. 3, pp. 685–692,2007.

[7] A. Fox, "PLAKSS - Psycolinguistische Analyse kindlicher Sprechst¨orungen," Swets & Zeitlinger, Frankfurt a.M., Germany, 2002.

[8] 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," Proceedings of International Conference on Pattern Recognition, vol. 4, pp. 274–277, 2006.

[9] A. Röbel, F. Villacencio and X. Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," Pattern Recognition Letters, vol.28, pp. 1343-1350, 2007.

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

[11] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, pp. 861-874, 2006.

Affiliation of the first Author

Lehrstuhl für Mustererkennung (Informatik 5),
Friedrich-Alexander Universität Erlangen Nürnberg,
Martensstr.3, 91058 Erlangen, Germany
cassia.valentini@i5.informatik.uni-erlangen.de