Automatic Detection of Sigmatism in Children

Cassia Valentini-Botinhao[1], *Sabine Degenkolb-Weyers*[2], *Andreas Maier*[3]
Elmar Nöth[3], *Ulrich Eysholdt*[4], *Tobias Bocklet*[3]

[1]The Centre for Speech Technology Research, University of Edinburgh, UK
[2]School of Speech Therapy, University Erlangen-Nuremberg, Germany
[3]Pattern Recognition Lab, University Erlangen-Nuremberg, Germany
[4]Department of Phoniatrics and Paedaudilogy, University-Clinics, Erlangen, Germany

C.Valentini-Botinhao@sms.ed.ac.uk, sabine.degenkolb-weyers@uk-erlangen.de
andreas.maier@informatik.uni-erlangen.de,noeth@informatik.uni-erlangen.de
ulrich.eysholdt@.uk-erlangen.de,tobias.bocklet@informatik.uni-erlangen.de

Abstract

We propose in this paper an automatic system to detect sigmatism from the speech signal. Sigmatism occurs when the tongue is positioned incorrectly during articulation of sibilant phones like /s/ and /z/. For our task we extracted various sets of features from speech: Mel frequency cepstral coefficients, energies in specific bandwidths of the spectral envelope, and the so-called supervectors, which are the parameters of an adapted speaker model. We then trained several classifiers on a speech database of German adults simulating three different types of sigmatism. Recognition results were calculated at a phone, word and speaker level for both the simulated database and for a database of pathological speakers. For the simulated database, we achieved recognition rates of up to 86%, 87% and 94% at a phone, word and speaker level. The best classifier was then integrated as part of a Java applet that allows patients to record their own speech, either by pronouncing isolated phones, a specific word or a list of words, and provides them with a feedback whether the sibilant phones are being correctly pronounced.

Index Terms: Gaussian Mixture Models, Support Vector Regression, Acoustic Analysis, Sigmatism

1. Introduction

Medical research can benefit from tools developed in areas like signal processing, machine learning and pattern recognition. Applications can range from diagnosis and treatment to interventional assistance. Our work here focuses on the choice of processing tools and their optimization for the classification of a speech disorder found mostly in children.

Sigmatism, also known as lisping, is a distortion of the phone /s/ and /z/. It is the most common type of speech disorder found in children. Depending on the tongue position, we can distinguish different types of sigmatism, such as interdental, dentalised and lateral. Interdental lisps happens when the tongue stretches out between the front teeth creating a /θ/ sound instead of a /s/ or /z/. It is the most common type of sigmatism [1]. Dentalised sigmatism occurs when the tongue is placed against the front teeth and air is pushed outwards. Lateral sigmatism occurs when the tongue touches the roof of the mouth, as if the talker was aiming to pronounce /l/, and the air is pushed outwards laterally. During the treatment of the disorder patients are assisted by speech therapists in weekly sessions, where they can have a professional feedback on how well they pronounce the target phones. At home patients would benefit from having a tool that could provide such a feedback as it is often hard for them to perceive differences in pronunciation.

Previous work on this task has focused on visualization tools to assist speech therapy [2] and only a few studies [3, 4] have been conducted on pathological speech data that could form the basis of a complete automatic evaluation system for this disorder. Collecting enough data of disordered speech to train a classifier is often a hard task as it involves finding a diverse collection of patients. Instead in this work we train our classification system with data we collected from speech therapist students simulating the disorder, a canonical representation of the disorder as seen by specialist in the area. To test the system we then use a small dataset of disordered speech.

In the following section we show how we collected datasets of simulated and pathological data, in Section 3 we present results of data analysis performed on the simulated dataset showing how disorder speech spectrum components differs from normal speech. In Section 4 we present the classification systems to be evaluated and Section 5 presents results of this evaluation with the two datasets at a phone, word and speaker level, followed by conclusions.

2. Data Collection

We collected two different datasets, one referred here as the *simulated database*, containing pathological speech data simulated by pupils of a speech therapy school. The other dataset contains real pathological speech data (*pathological speakers database*). The simulated database, as it has a larger amount of speakers, was used to train the classification system. The pathological speakers database was composed of speech from just a few individuals and therefore was used for testing purposes only.

The corpora of simulated data contains speech from 39 adults (37 female, 2 male). All participants had no speech disorder at the time of recording and all of them were pupils of the School of Speech Therapy of Erlangen, therefore able to simulate the three different types of sigmatism.

The pathological speakers database contains speech from six teenagers and one adult. Three of the teenagers had a certain type of sigmatism and were patients being treated by the School of Speech Therapy of Erlangen. The other three had no sigmatism or other speech disorder. The adult had dentalised sigmatism due to the use of a dental prosthesis. We were able

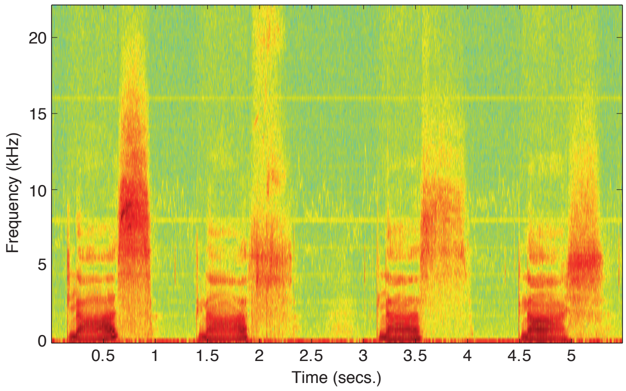

Figure 1: Spectogram of the German word "Glas" pronounced in four different manners: normal / interdental / dentalised / lateral.

to record both pathological and normal speech from him, once he removed the prosthesis.

All data were recorded with the use of a standard headset (dnt Call 4U Comfort) and sampled at $44.1\,$kHz with a $16\,$bit quantization. The databases were manually segmented at phone and word level.

Each individual uttered 16 German words that belonged to a test designed in the School of Speech Therapy of Erlangen. The test is mainly used to initially identify problems related to sibilant pronunciation. Each of the 16 words contains at least one realization of /s/ or /z/, in different positions.

3. Data Analysis

Figure 1 shows the spectogram of four realizations of the German word "Glas" from a particular speaker contained in the simulated database. We can see that due to coarticulation not only the /s/ segments differ from each other but also the neighboring phone.

For a more detailed acoustic analysis we processed the speech signals from the simulated database using a Hamming window of $25.6\,$ms duration and a frame rate of $100\,$Hz. From this segments we estimated the spectrum envelope through cepstral smoothing, using the true estimator technique [5] with a 2048 point DFT and cepstral smoothing order of 60.

Figure 2 shows the mean spectrum envelope for each pronunciation type, i.e. normal, interdental, dentalised and lateral, averaged over all speakers and over the /s/ and /z/ phones. The deviation within speakers was not greater than $6\,$dB ; within phones the deviation within speaker was not greater than $5\,$dB. The high peak located in the lower frequency range, below $300\,$Hz, in all figures corresponds to background noise present in all recordings, probably the noise of the computer in the recording room.

As we can see in Figure 2 the spectrum of the lateral sigmatism presents higher energy levels than the normal case in the region just below $5\,$kHz, as can also be seen in Figure 1. For frequencies higher than $5\,$kHz, normal speech presents higher levels than all lisping cases. Above $11\,$kHz the lateral curve stays below all other curves and falls faster at higher frequencies. Finally, we can see that interdental and dentalised spectrum practically coincide. This result is expected since the difference between these two misarticulations is only a slight change in the position of the tongue. Depending on the speaker, this difference was not always very evident.

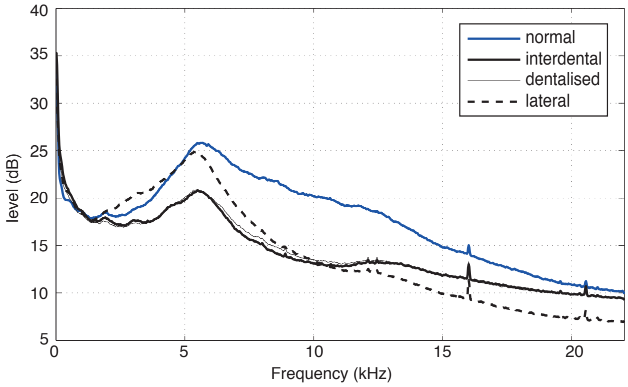

Figure 2: Mean spectrum envelope of /s/ for all speakers.

4. Classification System

Based on the results of the data analysis described in the previous section we choose to extract the following spectrum based type of features:

- **Energy**: a two dimensional feature vector containing the energy calculated form the spectral envelope in two different frequency bandwidths: 5-11 kHz and 11-20 kHz.
- **MFCC**: a 24 dimensional feature vector containing 12 static and 12 dynamic Mel-Frequency Cepstral Coefficients (MFCC). The Mel-filterbank consists of 22 filters. The dynamic features were calculated using a five frame window.
- **Supervector (SVector)**: a $M_k(1 + 2D)$-dimensional feature vector containing the concatenated parameters of a Maximum A Posteriori (MAP) adapted Gaussian Mixture Model (GMM). These parameters are the weights, mean vectors and diagonal elements of the covariance matrix of M_k Gaussian densities constructed with a D-dimensional MFCC feature vector. The GMM is MAP adapted from a Universal Background Model (UBM), which is a GMM trained with the data from all speakers and all types of realization, see [6] and [7]. The parameters of the MFCC frontend were the same as those previously described for the MFCC feature set;
- **Simplified Supervector (SSVector)**: a M_k-dimensional feature vector containing only the weight of the densities of the MAP adapted GMM.

For the classification task we compared the following classifiers, as implemented by the WEKA toolbox [8]:

- **NaiveBayes**: the class conditionals are modeled as a unimodal Gaussian distribution;
- **ViaRegression**: the class conditionals are approximated using a model tree [9];
- **SVM**: support vector machine classification with polynomial kernel, which generally showed better results;
- **AdaBoostM1**: AdaBoost classifier [10] using DecisionStump, a decision tree with only one node, as a weak classifier.

Feature	RR(%)	Rn(%)	Rp(%)	CL(%)	AUC
Energy	75.64	61.89	79.32	70.61	0.781
MFCCs	82.46	72.62	85.33	78.98	0.789
SVector	**86.00**	**77.19**	**88.93**	**83.06**	**0.830**
SSVector	82.79	71.79	86.46	79.14	0.791

Table 1: Phone level results: Energy feature set using ViaRegression and the MFCCs, SVector and SSVector using SVM.

Feature	RR(%)	Rn(%)	Rp(%)	CL(%)	AUC
SVector	**86.85**	68.42	**93.01**	**80.71**	0.807
SSVector	84.89	**70.51**	89.69	80.10	**0.863**

Table 2: Word level results: SVector feature set using SVM and SSVector using AdaBoostM1.

5. Results and Discussion

In this section we show classification results at a phone, word and speaker level. Each table compares the best performance obtained by the different feature sets in terms of absolute Recognition Rate on the whole set (RR), on the normal class (Rn) and the Pathological class (Rp) as well as the CLass-wise averaged recognition rate (CL) and the Area Under the receiver operating characteristics Curve (AUC). Training and testing were conducted on the simulated data in a Leave-One-Speaker-Out manner.

5.1. Simulated data

For the phone level results, we tested all the feature sets described in Section 4. However, for the word and speaker level, where the sibilants sounds were not segmented, only the Supervector and its simplified version were used.

5.1.1. Phone level

In total we had 2792 /s/ and /z/ phone recordings. The MFCC filterbank covered a bandwidth between 1 kHz and 16 kHz. To extract both SVector and SSVector we train a UBM using only /s/ and /z/ sounds with 16 densities to cover as many different types of pronunciation of the phones as possible.

Comparing the overall results for the two class problem in Table 1 we see that the SVector feature set had the best performance amongst all feature sets. The highest RR was of 86 %, for SVector using the SVM classifier. The second highest overall RR, 82.79 %, was also obtained by the SVM, but this time for the SSVector feature set, followed by 82.46 % and 75.64 % obtained by the SVM and the ViaRegression classifiers for the MFCCs and Energy feature set respectively. All three results are significantly different from the Supervector best result at a significance level of $p < 0.001$.

With only two static parameters the “Energy” feature set was able to achieve a classification rate slightly above 75 % compared to a 82 % obtained by the MFCC set of 24 parameters, that also included coefficients describing the dynamics.

5.1.2. Word level

The results on the word level were obtained using 2496 exemplars of recorded words. The UBM created for the Supervector extraction was trained using the speech sounds of all recorded words. We used more densities than in the phone level case – 128 – in order to cover other sounds and coarticulations present in the 16 word list.

In order to put more emphasis on the higher frequencies, to extract the MFCC features we used filterbanks that covered a bandwidth of 5-16 kHz. This bandwidth choice seems to perform better because it restricts the information that is modeled by the UBM to a more relevant frequency range.

The word level results are summarized in Table 2. The highest RR value of 86.85 % was found using the SVM classifier for the SVector case and 84.89 % was found using the AdaBoostM1 classifier for the SSVector feature set. At the word level SSVector and SVector results are significantly different only at a level of $p < 0.050$. Using the weights alone did not seem to have a big impact on the AUC and CL values either. In fact, it improved the correct classification for normal speech.

5.1.3. Speaker level

This set of experiments was performed on whole utterances. An utterance in this case is a list of all 16 words (with silence in between) uttered by a single speaker. All in all, 148 speakers' utterances, i.e., 37 speakers producing four different pronunciations each, were available. Two speakers were not used in this test because they were recorded in a different order.

Once again we present here classification results only for the Supervector feature types, i.e SVector and SSVector. The UBM was created using all recorded utterances, including words and silence. We used 128 densities, as this number gave better results in preliminary experiments. For the MFCC extraction we again used Mel-filterbanks that covered a bandwidth of 5-16 kHz.

The results at the speaker level is shown in Table 3. The highest RR values for each feature set - 92.56 % obtained by the SVM classifier for the Supervector and 93.91 % obtained by both SVM and AdaBoostM1 for the Simplified Supervector - are very similar and not significantly different from each other.

Feature	RR(%)	Rn(%)	Rp(%)	CL(%)	AUC
SVector	92.56	81.08	**96.39**	88.73	0.887
SSVector	**93.91**	**89.18**	95.49	**92.33**	**0.993**

Table 3: Speaker level results: both feature sets using SVM.

For the speaker level we additionally present results for the classification of sigmatism type in Table 4. The $\overline{\text{AUC}}$ values in Table 4 correspond to averages of AUC values taken over the four classes. The recognition rates for the pathological cases interdental, dentalised and lateral are presented as Rp1, Rp2 and Rp3 respectively. In the results of the four class problem the Rn value was highest among all class rates, since normal speech is the most distinctive class. Compared to recognition rates obtained for the pathological types, the lateral recognition rate Rp3 is the highest one. Interdental and dentalised sigmatism are the most difficult types to differentiate, and showed lower rates.

RR(%)	Rn(%)	Rp1(%)	Rp2(%)	Rp3(%)	AUC
62.83	89.18	48.64	40.54	72.97	0.844

Table 4: Speaker level results for the four class problem: SVectors using SVM.

5.2. Pathological speakers data

Using the best pairing of classifier/feature set for each level we show here the results of the classification task for the pathological dataset.

Class	Age	Phone	Word	Speaker
interdental	12	12/18	11/16	0/1
lateral	14	16/18	9/16	0/1
dentalised	18	17/18	14/16	1/1
normal	7	10/18	11/16	1/1
normal	11	12/18	8/16	1/1
normal	13	16/18	14/16	1/1
dentalised	27	16/18	14/16	1/1
normal	27	10/18	7/16	0/1
Total (%)		75.70	68.75	62.50

Table 5: Pathological speakers data classification results at all levels. For each speaker, 18 phones, 16 words and 1 list of words were tested. The results are in fractions and the overall result for each level is shown in percentage.

For all levels, we chose Supervector as the feature set. For the phone and word levels, the SVM classifier was chosen. For the speaker level, the AdaBoostM1 classifier was chosen. Both the classifier and the UBM were trained with speech from all 39 speakers of the simulated dataset. This configuration was incorporated in the Java applet for automatic sigmatism detection, illustrated in Figure 3.

Table 5 presents in each row the speakers available for testing the system and their classification results for each level. The results are shown in fractions of correctly classified items over tested units. The overall result for each level is given in the last row.

Overall compared to the recognition scores obtained with the simulated data the scores obtained with the real pathological data are smaller, but still a 75 % of recognition rate could be obtained at a phone level without the need to record disordered patients.

Table 5 also shows that the best results are obtained at the phone level, which is different from what was seen with the simulated data. The fact that we created the UBM only on adult speech might explains this result. For the phone level case, a UBM was created only with the /s/ and /z/ phones. The other levels' UBMs model coarticulation and other phones as well, which attenuates the effect of the mismatch between training and testing data.

6. Conclusions

In this paper we have proposed an automatic sigmatism detection tool for children. Our main goal was to classify the speech sounds as normal or pathological at the phone, word and speaker level, in order to assist patients treatment. We used simulated data, manually segmented into these units, for training and testing. We carried out leave-one-speaker-out experiments using different features and classifiers. At phone level we achieved a Recognition Rate (RR) of almost 86 %. At word level the RR value slightly increased to 87 % and at speaker level, we achieved a RR of 94 %. For the classification of sigmatism types, i.e. four class problem, we were able to achieve about 63 % of RR at speaker level. The lowest recognition rates were, as expected, from the dentalised and interdental classes, both below 50 %. The classification results on the pathological dataset were not as good because of the mismatch between training and test data, but still we achieved a 75 % recognition rate without the need of recording disordered speech data for training. We expect then that with a larger dataset, containing more children's speech, we can improve recognition rates on real pathological data. Using feature sets such as the Supervector one, which attempt to describe speaking style rather then phone characteristics, allowed us to process the speech signal without any sort of automatic segmentation step. For an improved version of the system though, the use of an automatic speech/silence segmentation module would no doubt be beneficial. Once we have collected a larger database of pathological data, we have plans to make the system publicly available in the internet.

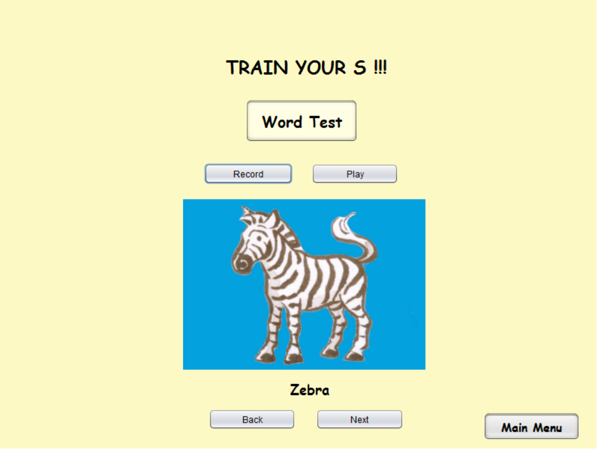

Figure 3: Screenshot of the main page of the Java applet

7. References

[1] M. Weinrich and H. Zehner, *Phonetiche und Phonologische Störungen bein Kindern*. Heidelberg, Germany: Springer, 2003.

[2] K. Grauwinkel and S. Fagel, "Visualization of internal articulator dynamics for use in speech therapy for children with sigmatismus interdentalis," in *Int. Conf. on Auditory-Visual Speech Processing*, 2007, paper 32.

[3] M. Akagi, N. Suzuki, K. Hayashi, and H. Saito, "Perception of lateral misarticulation and its physical correlates," *Folia Phoniatr Logop*, vol. 53, no. 6, pp. 291–307, 2001.

[4] Z. Benselama, M. Guerti, and M. Bencherif, "Arabic speech pathology therapy computer aided system," *J. of Computer Science*, vol. 3, no. 9, pp. 685–692, 2007.

[5] A. Röbel and F. V. X. Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Recogn. Lett.*, vol. 28, no. 11, pp. 1343–1350, 2007.

[6] T. Bocklet, T. Haderlein, F. Hönig, F. Rosanowski, and E. Nöth, "Evaluation and assessment of speech intelligibility on pathological voices based upon acoustic speaker models," in *Proc. of the 3rd Advanced Voice Function Assessment Int. Workshop*, 2009, pp. 89–92.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, pp. 19–41.

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

[9] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. Witten, "Using model trees for classification," *Machine Learning*, vol. 32, no. 1, pp. 63–76, 1998.

[10] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Thirteenth Int. Conf. on Machine Learning*. San Francisco: Morgan Kaufmann, 1996, pp. 148–156.