ANALYSIS OF HYPERNASAL SPEECH IN CHILDREN WITH CLEFT LIP AND PALATE

Andreas Maier[1,2], Alexander Reuß[1], Christian Hacker[1], Maria Schuster[2], and Elmar Nöth[1]

[1] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5)
Martensstraße 3, 91058 Erlangen, Germany
[2] Universität Erlangen-Nürnberg, Abteilung für Phoniatrie und Pädaudiologie
Bohlenplatz 21, 91054 Erlangen, Germany
andreas.maier@informatik.uni-erlangen.de

Abstract. In children with cleft lip and palate speech disorders appear often. One major disorder amongst them is hypernasality. This is the first study which shows that it is possible to automatically detect hypernasality in connected speech without any invasive means. Therefore, we investigated MFCCs and pronunciation features. The pronunciation features are computed from phoneme confusion probabilities. Furthermore, we examine frame level features based on the Teager Energy operator. The classification of hypernasal speech is performed with up to 66.6 % (CL) and 86.9 % (RR) on word level. On frame level rates of 62.3 % (CL) and 90.3 % (RR) are reached.

1 INTRODUCTION

In clinical practice it is desirable to objectively quantify the severity of speech disorders by non-invasive means. The state-of-the-art techniques to measure hypernasality today are quite invasive since the patients' nasal and oral airflows have to be measured. This is usually done with devices like a Nasometer [1] which is placed between the mouth and the nose in order to separate both airflows. This procedure is complicated—especially with children.

Non-invasive methods exist [2, 3], however, their application demands a lot of manual preprocessing since these methods can only be applied to sustained vowels or consonant-vowel combinations. In the literature the segmentation is usually done manually which costs a lot of time and effort. In order to close this diagnostic gap we want to investigate, if a fully automatic evaluation system can be applied for such a task. Therefore, an automatic speech recognition system is used to segment the audio data into words. To improve the automatic segmentation the transliteration of the speech data was supplied to the speech recognition system. In a next step, of course, we will replace the manual transliteration with an automatic speech recognition system. To train and evaluate an automatic classifier, a speech therapist labeled all words either as "hypernasal" or "normal".

This research is being integrated into our automatic speech evaluation platform [4] environment, which is a web application to analyze and evaluate various speech disorders. At the moment our system can already judge the speaker's intelligibility. To achieve this, it uses a speech recognizer and calculates either word accuracy (WA) or word recognition rate (WR) as a measure for intelligibility [5].

The aim of this work is to take our system one step further. It is our intention to enable the ability to detect specific speech disorders. In the following sections, we describe facial clefts and analyze different features for the detection of hypernasality.

2 CLEFT LIP AND PALATE

Cleft lip and palate (CLP) is a frequent congenital alteration of the face with a prevalence of about one in 750 to 900 births [6]. Today, the visible effects of CLP can be surgically lessened. If a grown up patient has been treated well, it is hardly noticeable, that he had a facial cleft as a child. Apart from surgical interventions, the patient also receives speech therapy. This is necessary, because the alteration can have a major impact on the patient's vocal tract and can lead to various speech disorders. The most common is hypernasality which is caused by enhanced nasal air emissions. The first formant is less distinct while antiresonances and nasal formants appear [7]. As a consequence, vowels are perceived with a characteristic nasal sound. However, speakers also have problems with other phonemes: Fricatives can not be pronounced correctly and plosives are weakened [8].

During the speech therapy of a patient with hypernasality due to cleft lip and palate, an automatic system to detect hypernasal speech would be very useful because it can make the treatment easier by providing a way to keep track of the patient's progress.

3 CLASSIFICATION SYSTEM

All experiments use a Gaussian mixture model (GMM) classifier according to the following decision rule:

$$k = \underset{\kappa}{\operatorname{argmax}}\ P(\Omega_\kappa) \cdot P(\boldsymbol{c} \mid \Omega_\kappa) \tag{1}$$

Here, κ denotes a class, $\boldsymbol{c}$ is our feature vector and Ω_κ is the event that the current observation belongs to class κ. In our case, there are only two classes: $\kappa \in \{\text{nasal}, \text{normal}\}$.

The probability $P(\boldsymbol{c} \mid \Omega_\kappa)$ is approximated by a mixture of M Gaussian densities $\mathcal{N}(\boldsymbol{c}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ with mean vectors $\boldsymbol{\mu}_m$ and covariance matrices $\boldsymbol{\Sigma}_m$:

$$\begin{aligned} P(\boldsymbol{c} \mid \Omega_\kappa) &\approx \sum_{m=1}^{M} a_m \mathcal{N}(\boldsymbol{c}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \\ \text{with } \sum_{m=1}^{M} a_m &= 1 \end{aligned} \tag{2}$$

Our classifier is trained by calculating a_m, $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ by means of the expectation maximization algorithm [9].

3.1 Word dependent a-priori probabilities

We estimate the prior probabilities $P(\Omega_{\text{nasal}})$ by calculating the frequency of hypernasality for every word w from the training set. If w is never marked as hypernasal in our training set, this would lead to a zero probability which means a record of w can never be classified as hypernasal. Therefore, we interpolate by equally distributing the probability mass of the words that are marked as hypernasal less than two times:

$$P_w(\Omega_{\text{nasal}}) \approx \begin{cases} \dfrac{\#\text{nasal}_w}{\#\text{total}_w} & \text{if } \#\text{nasal}_w > 1 \\[2ex] \dfrac{\sum\limits_{z \in S} \#\text{nasal}_z}{\sum\limits_{z \in S} \#\text{total}_z} & \text{else} \end{cases}$$

S denotes the subset of words that were marked less than two times. $\#\text{nasal}_w$ is the number of times the word w was marked and $\#\text{total}_w$ is the number of times w occurs.

3.2 Pronunciation Features

Pronunciation features, as described in [10], were designed to rate a speaker's pronunciation. They are used for measuring the progress when learning a foreign language. In this work, we study these features' applicability to the detection of hypernasal speech. More precisely, we only analyze a subset of these features that is based on phoneme confusion probabilities on word level. To calculate these phoneme confusion features we compare the result of the forced alignment of every word to the result of a phoneme recognizer. The phoneme recognizer uses semi continuous hidden Markov models and a 4-gram language model. It is based on MFCCs calculated every 10 ms with a frame size of 16 ms. From these informations phoneme confusion matrices $\boldsymbol{C}$ are built. They contain for every pair of phonemes a, b the probability that a was detected by the recognizer when there should be b according to the forced alignment, i.e.,

$$\boldsymbol{C}_{a,b} = P(a \mid b) \tag{3}$$

From the training set, we calculate two confusion matrices: one for the hypernasal data and one for the normal data. We need to recalculate these matrices in every iteration of the LOO evaluation (cf. Sect. 4) because, in order to obtain representative results, the current test speaker may not be involved in the generation of our phoneme confusion model. The quotient Q is calculated for every frame:

$$Q = \frac{P_{\text{nasal}}(a \mid b)}{P_{\text{normal}}(a \mid b)} \tag{4}$$

From these frame-wise results, we calculate the following features for the word level:

- PC1: mean of Q
- PC2: maximum of Q
- PC3: minimum of Q
- PC4: scatter of Q
- PC5: median of Q

3.3 Cepstral Features

Furthermore, we investigate cepstral features, more accurately Mel frequency cepstral coefficients (MFCCs). These features are calculated frame wise and consist of the short time signal energy, 11 MFCCs and 12 dynamic features in a context of ± 2 frames i.e. 56 ms. Using these features, we train a frame based classifier to analyze hypernasal vs. normal speech. However, the expert annotations have all been performed on the word level. Thus, the expert labels have to be mapped onto the frame level. In a first approach, we simply label every frame with the respective word label; then a frame based classifier is trained. As it is known for every frame, which word it was taken from, we can still use the word based a-priori probabilities (cf. Section 3.1) as we did before.

To compare the classification results with previous investigations with word based pronunciation features, the evaluation is again performed on the word level. This means, that in the test phase a decision for the class of a word is derived from all the classification results for its frames. There are several ways of making a word level decision from the frame level (e.g. mean or median) but the best results were achieved using the maximum frame wise classifier score as the classification hypotheses for the whole word.

As mentioned above, during training all frames are labeled the same as the word they belong to. However, if a word is labeled "hypernasal" that does not mean that *every* part of this word is hypernasal. So we got normal frames labeled as hypernasal in our training procedure. We tackle that issue using a bootstrapping algorithm (similar to [11]). We train our frame wise classifier just like we did before. Then we classify the training data and relabel the frames of the hypernasal words with the hypothesis of the frame based classifier. This process is iterated a fixed number of times. We choose two iterations per word as preliminary experiments showed that more iterations do not yield further improvements.

3.4 Teager Energy Operator

The next feature we evaluate is the Teager Energy operator (TEO) [2]. It is defined as:

$$\psi[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{5}$$

$x(n)$ denotes the time domain audio signal. The TEO's output is called the Teager Energy profile (TEP).

As already described in [2], the TEP can be used to detect hypernasal speech because it is sensitive to composite signals. When normal speech is lowpass filtered in a way that the maximum frequency f_{lowpass} is somewhere between the first and the second formant, the resulting signal mainly consists of the first formant. However, doing the same with hypernasal speech results in a composite signal due to the strong anti formants. If we now compare the lowpass filtered TEP to the TEP of the same signal that was bandpass filtered around the first formant we should see more difference in case of a hypernasal signal. We measure that difference with the correlation coefficient of these TEPs. In the following, the bandpass filter covers the frequency range ±100 Hz around the first formant estimated using PRAAT [12].

We got the following experimental setup: We use the correlation between both TEPs as a feature and calculate it for every frame. Then, we apply a phoneme recognizer to determine the phoneme, the frame belongs to. The classifier is trained per vowel with the features from all frames of the training words that were associated with that vowel. Afterwards, these classifiers are tested with the frames from the words of the test speaker that were assigned to the respective vowel.

4 Evaluation

As our data set is rather small we use leave-one-speaker-out (LOO) evaluation to rate our classifiers. There are much more normal words than hypernasal words, so recognition rate (RR) is not very meaningful. Therefore, we calculate the mean recognition rate per class (CL) as well. It is the average of the recalls for the classes "hypernasal" and "normal".

$$\mathrm{CL} = 0.5(\mathrm{REC}_{\mathrm{nasal}} + \mathrm{REC}_{\mathrm{normal}}) \tag{6}$$

where $\mathrm{REC}_{\mathrm{nasal}}$ is the recall of the class "hypernasal" and $\mathrm{REC}_{\mathrm{normal}}$ the recall of the class "normal".

5 DATA

The data we use consists of recordings of 3 girls and 10 boys (5 to 11 years old) with cleft lip and palate recorded during the PLAKSS test [13]. Pictures are shown to the children which illustrate the words the children should speak. There are 99 different words, that contain all phonemes of the German language at 3 different positions in the word: at the beginning, at the end, and in the center. The single words were extracted from the recordings using forced alignment and were labeled as "hypernasal" or "normal" by an experienced speech therapist.

Since some children skipped some words and others had a quite low intelligibility some words could not be properly segmented by the forced alignment. These words had to be excluded from the data. In total we got 771 words. 683 of them are labeled "normal" and 88 are labeled "hypernasal". As some phonemes are more likely to be mispronounced due to hypernasality than others, the probability $P_w(\Omega_{\mathrm{nasal}})$ for being marked hypernasal is different for every word.

In order to keep the data as realistic as possible slight errors in the forced alignment were kept in the database (some samples were cut off at the beginning or at the end of the word). The audio files were stored with a sampling rate of 16 kHz and quantized with 16 bit per sample.

Table 1. Results on word level obtained with different features. MFCCs and pronunciation features yield feasible results.

FEATURE	RR	CL
PC1	**86.9 %**	57.7 %
PC2	86.4 %	52.4 %
PC3	82.7 %	**64.1 %**
PC4	86.8 %	56.3 %
PC5	86.1 %	57,7 %

(a) pronunciation features

DENSITIES	RR	CL
2	70.6 %	**65.1 %**
5	74.7 %	64.4 %
10	**75.5 %**	64.9 %
15	**75.5 %**	63.9 %

(b) MFCCs (no bootstrap)

DENSITIES	RR	CL
2	82.1 %	**66.6 %**
5	82.3 %	64.8 %
10	82.2 %	64.7 %
15	**83.1 %**	65.2 %

(c) MFCCs (bootstrap)

6 RESULTS

LOO evaluation of the single pronunciation features with our Gaussian classifier (we choose $M = 1$, more densities have shown to decrease the rates in this case) leads to the results shown in Table 1 (a). The idea behind choosing these features is that hypernasal speakers have problems pronouncing specific phonemes (plosives, some fricatives and some vowels). Therefore, the phonemes a recognizer does not identify properly should, to some extend, be similar for the nasal speakers. The class-wise recognition rate of up to 64.1 % verifies our assumption. As our training sets are rather small the confusion probabilities $P(a|b)$ can not be calculated very exactly. Therefore, we expect even better results in future experiments with more training data.

Testing the word level classification system based on frame-wise MFCCs as described in Sect. 3.3 leads to the results shown in Table 1 (b) and Table 1 (c). It can be seen that bootstrapping slightly improves the class wise recognition rate while considerably improving the total recognition rate.

The results of the frame-wise classification of the Teager Energy correlation feature as described in Sect. 3.4 were also promising. The formant frequencies for the bandpass were extracted with “Praat” automatically and a bandpass with a bandwidth of 200 Hz around the first formant was performed for the one TEP. For the other TEP we run 4 series of tests with lowpass cutoff frequencies $f_{\mathrm{lowpass}} =$ 1000 Hz, 1300 Hz, 1600 Hz, and 1900 Hz. Then, the correlation between both TEPs was determined and fed to the classification procedure. Table 2 shows the best results of for each of the vowels which appeared in the test data (vowels in SAMPA notation).

The results show, that the TEO can be used to classify hypernasal speech of children with cleft lip and palate. However, the rates are not as good as [2] might let expect. We see two reasons for this: first, we have no phoneme level annotation (a similar problem to what we discussed before regarding the MFCCs), second this concept does not work as good with children as their formants are harder to find (more detection errors) than those of adults. Due to the difficulties in the determination of the cutoff frequencies and that the TEP is only suitable for vowels we did not study their performance further on word level, yet.

Table 2. Results for frame wise Teager Energy features for different vowels and best cutoff frequency.

VOWEL	RR	CL	CUTOFF f_{lowpass}
/9/	**90.3 %**	62.5 %	1900 Hz
/a/	80.0 %	59.2 %	1000 Hz
/i:/	84.8 %	60.1 %	1900 Hz
/o:/	87.0 %	**63.2 %**	1000 Hz
/O/	88.2 %	55.6 %	1900 Hz
/u:/	90.0 %	55.6 %	1000 Hz
/U/	89.4 %	56.1 %	1900 Hz

7 DISCUSSION AND OUTLOOK

In the results section encouraging results for the classification of hypernasality in children's speech from automatically segmented audio data were presented: The class-wise recognition rate CL reaches up to 66.6 % and the recognition rates RR is in one case even 89.4 (53.4 % CL). We explain this effect with the fact that nasality detection in children's speech is more difficult than in adults' speech. Misdetection of normal speech as hypernasal speech, however, happened rarely in the best classifiers.

We still see some room for improvement in our future work. As our data set is relatively small, classification results could be greatly enhanced by using more data (further recordings were already performed). This will help estimating the prior probabilities for the GMM classifier and the confusion matrices for the pronunciation features. The combination of multiple features will also improve the performance of the classification.

Another possibility to improve the results is the usage of a phoneme level annotation. This will be a sensible step, because only some phones of a hypernasal word show nasal characteristics. The other phones of the realization might still be perceived as normal. The recognition rates of the MFCCs and the TEO features could benefit from it. Moreover, we want to investigate whether the TEP features can be enhanced in order to be applicable on word or speaker level, since their classification rates look quite promising. We expect to be able to use the techniques presented here soon in clinical practice.

8 SUMMARY

In this study we could show that the classification of hypernasality in children's speech on automatically segmented data is possible. We described the evaluation of several features regarding their suitability to classify hypernasal speech of children with cleft lip and palate. On word level, class-wise recognition rates of up to 66.6 % and global recognition rates of 86.9 % were achieved. First, we extracted pronunciation features based on phoneme confusion statistics. With these, we reached a CL of up to 64.1 % and a RR of 86.9 %. MFCC features were best in CL. We extracted them frame-wise and derived a word level decision from that. With a bootstrapping approach, we improved the annotation which led to rates of up to 66.6 % CL and 83.1 % RR. Finally, we studied the TEO on frame level. It was tested using separate classifiers for the frames belonging to different vowels which were identified using a simple phoneme recognizer. The results showed, that the TEO's performance varied for different phonemes and that it does not work as well in our scenario as in preceeding works with adult speakers and manually segmented consonant-vowel and consonant-vowel-consonant clusters.

References

1. Kay Elemetrics Corporation, New York, USA, *Instruction manual of the nasometer Model 6200-3, IBM PC Version*, 1994.
2. D. Cairns, J.H.L. Hansen, and J. Riski, "A Noninvasive Technique for Detecting Hypernasal Speech Using a Nonlinear Operator," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 1, pp. 35–45, 1996.
3. A. Zečević, *Ein sprachgestütztes Trainingssystem zur Evaluierung der Nasalitiät*, Ph.D. thesis, Universität Mannheim, Germany, 2002.
4. 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.
5. E. Nöth, A. Maier, T. Haderlein, K. Riedhammer, F. Rosanowski, and M. Schuster, "Automatic Evaluation of Pathologic Speech — from Research to Routine Clinical Use," in *10th International Conf. on Text, Speech and Dialogue (TSD)*, V. Matoušek and P. Mautner, Eds., Berlin, Heidelberg, New York, 2007, vol. 4629 of *Lecture Notes in Artificial Intelligence*, pp. 294–301, Springer.
6. M. Tolarova and J. Cervenka, "Classification and birth prevalence of orofacial clefts," *Am. J. Med. Genet.*, vol. 75, no. 2, pp. 126–137, 1998.
7. G. Fant, "Nasal Sounds and Nasalization," in *Acoustic Theory of Speech Production*, The Hague, The Netherlands, 1960, Mouton.
8. O. Braun, *Sprachstörungen bei Kindern und Jugendlichen*, Kohlhammer, Stuttgart, Germany, 2002.
9. T. Bocklet, A. Maier, and E. Nöth, "Text-independent Speaker Identification using Temporal Patterns," in *10th International Conf. on Text, Speech and Dialogue (TSD)*, V. Matoušek and P. Mautner, Eds., Berlin, Heidelberg, New York, 2007, vol. 4629 of *Lecture Notes in Artificial Intelligence*, pp. 318–325, Springer.
10. C. Hacker, T. Cincarek, R. Gruhn, S. Steidl, E. Nöth, and H. Niemann, "Pronunciation feature extraction," in *Pattern Recognition, 27th DAGM Symposium, August 30 - September 2005, Vienna, Austria, Proceedings*, Berlin, Heidelberg, Germany, 2005, Lecture Notes in Computer Science, pp. 141–148, Springer.
11. S.P. Abney, "Bootstrapping," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, San Francisco, USA, 2002, pp. 360–367.
12. P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341,345, 2001.
13. A. V. Fox, *PLAKSS - Psycholinguistische Analyse kindlicher Sprechstörungen*, Swets & Zeitlinger, Frankfurt a.M., now available from: Harcourt Test Services GmbH Frankfurt a.M., Germany, 2002.