Hypernasality in Speech of Children with Cleft Lip and Palate: Automatic Evaluation

Andreas Maier[1,2], Maria Schuster[1], Tino Haderlein[1,2], Elmar Nöth[2]

[1] University Erlangen-Nuremberg, Division of Phoniatrics and Pedaudiology, Bohlenplatz 21, 91054 Erlangen

[2] University Erlangen-Nuremberg, Chair of Pattern Recognition, Martensstr. 3, 91058 Erlangen

E-mail: andreas.maier@cs.fau.de

Abstract

We propose a novel approach for the automatic assessment of hypernasality in the speech of children with cleft lip and palate. Our system is based on short-time features as well as on large-scale features. It is compared to the perceptive evaluation of two experienced speech therapists. While the correlation between the scores of the speech therapists is 0.80 the correlation between the composite annotation of both raters and our proposed system is 0.81. The system operates on experts' level.

1 Introduction

Communication disorders pose a major challenge to society in the 21st century [1]. Speech of children with cleft lip and palate (CLP) may contain communication disorders – or more precisely speech disorders. The main feature of their speech is hypernasality. Other disorders as backing of the point of articulation and weakening of plosives may also occur.

Hypernasality is caused by additional nasal air emission in vowels and/or consonants due to insufficient velopharyngeal closure. Typical acoustic features of nasal vowels are antiformants which appear below the first formant (cf. Figure 1). Nasalized consonants typically hold additional noise in high frequencies caused by the enhanced nasal air stream [2].

Current state-of-the-art methods for the automatic evaluation of nasality are either invasive or may not be tolerated by children easily, i.e., a device has to be attached to or inserted into the body in order to measure the nasality [3], or they are limited to the analysis of vowels or simple consonant-vowel combinations [4]. The scope of this paper is to show that automatic speech recognition techniques and automatic feature extraction are able to assess hypernasality in connected speech of children.

2 Material

Speech data of 32 children with CLP were recorded. All children were recorded during the regular out-patient examination in the Cleft Lip and Palate Centre of the University Hospital Erlangen. Informed consent was obtained from all parents prior to the examination. The children showed a broad range in hypernasality ranging from almost no nasality to strong nasal air emission.

All children performed the "Psycho-Linguistische Analyse kindlicher Sprechstörungen" (PLAKSS) [5], a standard speech test for German children. The test contains all German phonemes at different positions within a vocabulary of 99 words.

3 Methods

The data was analysed by two speech therapists independently. Both annotated hypernasality on phone level. Evaluation on speaker level was

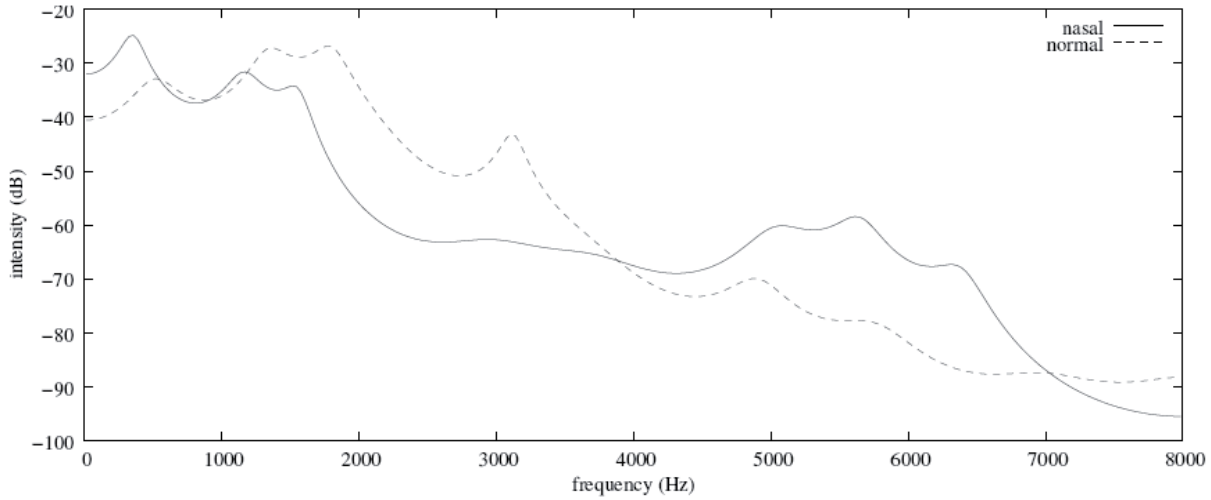

Figure 1: LPC *model spectra with 20 coefficients of a nasal and a non-nasal vowel /a:/ in the phonetic context /ha:s@/ produced by two CLP children. A nasal formant (~300–500 Hz) which is stronger than the first formant (~1100 – 1300 Hz) appears in the nasal realization.*

achieved by computation of the percentage of marked words. Then Pearson's correlation between both raters was determined to measure the agreement [6].

The correlation was also computed between the automatic nasality classification system and the subjective evaluation of the raters. However, a "golden standard" how to train the classification system had to be determined first. We chose to mark a phone as "nasalized" if both raters agreed on their decision. In case of disagreement, the phone was marked as "normal".

Figure 2 shows the setup of the evaluation system. First the speech data is segmented into different evaluation levels using our speech recognition system [7]. In order to determine whether a phone or word was realized as hypernasalized, characteristic features have to be extracted. This is performed on different levels in this work.

Starting from a very small time interval of 16 ms with short-time Mel Frequency Cepstral Coefficients (MFCCs) [8], we extract features on frame, phone, word, and speaker level. On phone level we evaluate so-called "Goodness of Pronunciation" (GOP) features, phone confusion features [9], and Teager Energy Profiles (TEPs) [4]. On word level more phone confusion features [9] and prosodic features [10] are evaluated. On speaker level this is extended with recognition accuracy features [11] and coordinates obtained from a Sammon map [12]. Furthermore, the features of the respective lower level are used to compute functionals such as mean, minimum, maximum, and standard deviation, and supplied to the next higher level. An overview on the different features is presented in Table 1. The training of the different classifiers was performed using the WEKA [13] implementation in leave-one-speaker-out (LOO) conditions.

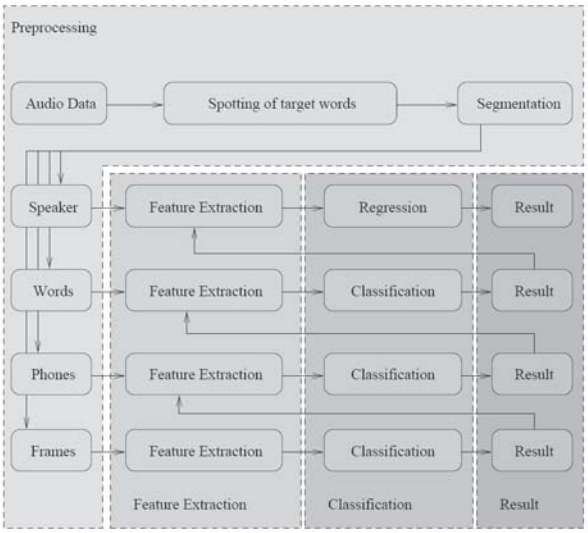

Figure 2: *Setup of the evaluation system: Features are extracted on each segmentation level. Then, a classification whether the observation was nasal or not is performed. The result is then lifted to the next higher evaluation level using functionals.*

Table 1: *Overview on the different features used.*

Level	#	Description	Ref.
Speaker	2	Accuracy of speech recognition (word correctness and accuracy)	[14]
speaker	2	Coordinates on a 2-D Sammon map	[12]
speaker	2	Coordinates on a 3-D Sammon map	[12]
word	37	Features based on the energy, the F0, pauses, and duration to model the prosody of the speaker	[10]
word	7	Pronunciation features to score the correctness of the current word	[15]
phone	6	Features to score the correctness of the pronunciation of the current phone	[16]
phone	1	Teager Energy Profile to detect nasality in vowels	[4]
frame	24	Mel Frequency Cepstrum Coefficients	[17]

On speaker level all computed features are put into a prediction system in order to determine a percentage of nasalized words via Multiple Correlation / Regression Analysis [18]. This is then compared to the subjective evaluation.

4 Results

The inter-rater agreement of the perceptive evaluation is high. Table 2 shows the results of the word level evaluation. The correlation between both raters is significant with $r = 0.80$ ($p < 0.01$).

The recognition results on frame, phone, and word level were joined in order to estimate the percentage of hypernasal words on speaker level. The agreement between the automatic system and the human evaluation is very good. The comparison of the automatic system to the perceptive ratings is in the same range with 0.81 ($p < 0.01$).

Table 2: *Confusion matrix for both speech therapists who rated the criterion "nasality": Both raters show high agreement on the criterion "non-nasal". The agreement in the "nasal" case is rather low.*

nasality	Nasal (rater 1)	non-nasal (rater 1)
nasal (rater 2)	127	203
non-nasal (rater 2)	152	2499

5 Summary

Hypernasality in speech of children was successfully evaluated using a system based on state-of-the-art features in automatic pronunciation scoring. A reliable prediction of hypernasality could be achieved which is in the same range as the evaluation of experts on speaker level.

6 Acknowledgments

This work was funded by the German Research Council (Deutsche Forschungsgemeinschaft) under grant SCHU2320/1-1. We would like to thank Andrea Schädel and Dorothee Großmann for the expert's annotation of the data.

References

[1] R. Ruben. "Redefining the survival of the fittest: communication disorders in the 21st century". Laryngoscope, Vol. 110, No. 2, pp. 241-245, 2000.
[2] G. Fant. "Nasal Sounds and Nasalization". In: Acoustic Theory of Speech Production, Mouton, The Hague, The Netherlands, 1960.
[3] Instruction manual of the nasometer Model 6200-3, IBM PC Version. Kay Elemetrics Corporation, New York, USA, 1994.
[4] D. Cairns, J. 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.
[5] A. Fox. "PLAKSS – Psycholinguistische Analyse kindlicher Sprechstörungen". Swets & Zeitlinger, Frankfurt a.M., Germany, now available from Harcourt Test Services GmbH, Germany, 2002.
[6] K. Pearson. "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia". Philosophical Transactions of the Royal Society of London, Vol. 187, pp. 253–318, 1896.
[7] G. Stemmer. Modeling Variability in Speech Recognition. Logos Verlag, Berlin, Germany, 2005.
[8] H. Niemann. Klassifikation von Mustern. available online, 2nd Ed., 2003. http://www5.informatik.uni-erlangen.de/Personen/niemann/klassifikationvon-mustern/m00links.html; last visited 02/12/2008.
[9] C. Hacker, T. Cincarek, R. Gruhn, S. Steidl, E. Nöth, and H. Niemann. "Pronunciation Feature Extraction". In: G. Kropatsch, R. Sablatnig, and A. Hanbury, Eds., Pattern Recognition, 27th DAGM Symposium, Vienna, Austria, Proceedings, pp. 141–148, Springer, Berlin, Heidelberg, Germany, 2005.
[10] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. "The Prosody Module". In: W. Wahlster, Ed., Verbmobil: Foundations of Speech-to-Speech Translation, pp. 106–121, Springer, New York, Berlin, 2000.
[11] T. Haderlein. Automatic Evaluation of Tracheoesophageal Substitute Voices. Logos Verlag, Berlin, Germany, 2007.
[12] T. Haderlein, D. Zorn, S. Steidl, E. Nöth, M. Shozakai, and M. Schuster. "Visualization of Voice Disorders Using the Sammon Transform". In: P. Sojka, I. Kopecek, and K. Pala, Eds., 9th International Conf. on Text, Speech and Dialogue (TSD), pp. 589–596, Springer, Berlin, Heidelberg, New York, 2006.
[13] I. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Fransisco, USA, 2nd Ed., 2005.
[14] 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.
[15] C. Hacker, T. Cincarek, A. Maier, A. Heßler, and E. Nöth, .Boosting of Prosodic and Pronunciation Features to Detect Mispronunciations of Non-Native Children,. in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. Hawaii, USA: IEEE Computer Society Press, 2007, pp. 197-200.
[16] T. Cincarek. Pronunciation Scoring for Non-Native Speech, Diplomarbeit, Chair of Pattern Recognition, University Erlangen-Nuremberg, Erlangen, Germany, 2004.
[17] S. Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences". *IEEE Trans. on Acoustics, Speech, and Signal Processing (ASSP)*, vol. 28, no. 4, pp. 357-366, 1980.
[18] J. Cohen and P. Cohen. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1983.