Towards the Automatic Evaluation of Dysarthric Speech

Andreas Maier, Elmar Nöth, Frank Rosanowski, Christa Sous-Kulke, Wilfried Schupp, University of Erlangen-Nuremberg, University Clinic Erlangen, m&i Fachklinik Herzogenaurach, Germany

Abstract

After a stroke, the quality of the speech in patients is often reduced. This is usually caused by a deficit of the motor abilities of the vocal tract. The result is slurred speech. In the various patients, however, very different forms can appear. In the course of therapy, evaluation of the speech quality is required to determine the success of the treatment. At the moment, this assessment is performed only perceptually. This form of assessment is subject to strong intra- and inter-individual variation. Therefore, an "objective" assessment is not guaranteed. In this study we present a rater-independent method for the speech disorders in dysarthria. We use methods of automatic speech recognition. The idea is to determine the speech intelligibility - the main outcome parameter of speech - automatically by an automatic speech recognizer. Between the criterion "intelligibility" and the recognition rate of the automatic system a correlation of -0.89 was obtained in a first preliminary study.

Contact: Andreas Maier, Andreas.Maier@cs.fau.de, 09131/8527872

1. Introduction

There is no objective, validated, automated procedure for the determination of the speech intelligibility in patients with dysarthria. The perceptual assessment of the intelligibility by speech therapists is not objective and, therefore, subject to inter- and intra-individual variation. In particular, experience is a crucial factor [1]. In order to obtain a more reliable assessment, patients are often evaluated by an expert committee or panel. However, this is usually performed only for clinical studies and research, because a lot of time and effort are required. In this paper we present the use of an automatic speech recognition system to evaluate the intelligibility. Furthermore, we use automatic prosodic features which are also extracted from the speech signal and compare them with a number of other perceptual criteria.

2. Method

The speech data were recorded over the Internet with our “Program for Evaluation and Analysis of All Kinds of Speech” (PEAKS) [2]. PEAKS runs in any Internet browser and is based on Java technology which allows platform-independent use. The data are transmitted to our server and evaluated centrally (cf. Figure 1). For this we use a speech recognition system which is based on Hidden Markov models (HMMs). As training data solely normal speakers were used [3]. The training population covered most dialectal regions of Germany. Result of the analysis is the number of correctly recognized word with respect to the reference.

$$WC = (C / R) * 100\ \% \qquad (1)$$

C denotes the number of correctly recognized words, and R the number of words in the reference.

Furthermore, automatic prosodic features which model energy, fundamental frequency, length of voiced and voiceless segments, jitter, and shimmer were investigated. The results of the analysis are available briefly after recording. To compare the human evaluation and the automatic one, five speech therapists with at least five years of experience rated the criteria “intelligibility”, “roughness”, and “prosody”. They were performing a five-point scale assessment for each criterion and the average per patient was computed in order to obtain rater-independent scores. The agreement between the human and the automatic evaluation was determined as the correlation according to Pearson [4].

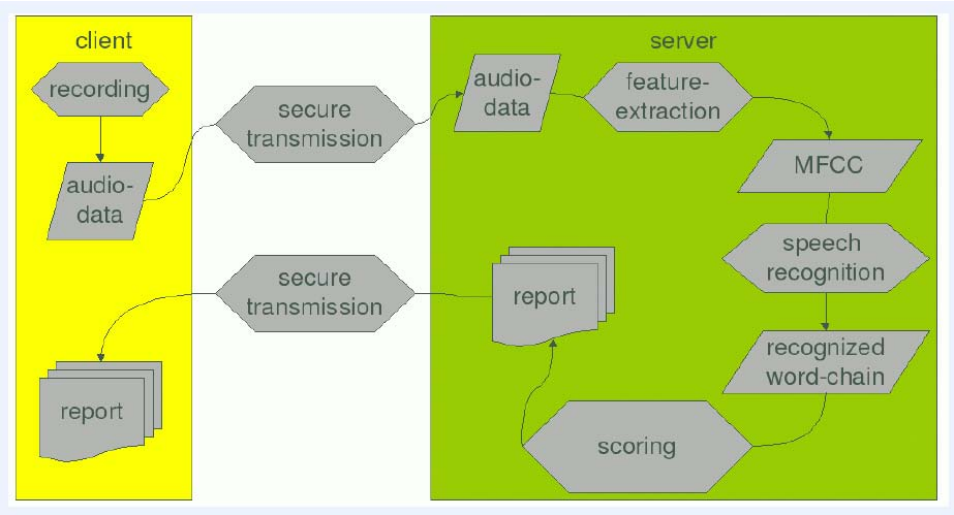

Figure 1: Diagram of the client-server architecture of PEAKS

3. Patients

For this study, nine patients with dysarthria were recorded. The patients were 39 to 76 years old. Depending on the severity of the dysarthria, the treatment can take 3 to 18 weeks. The average duration is 5 weeks.

The data consisted of read speech of a standard text (“The North Wind and the Sun”). Overall, the text contains 108 words, of which 71 are disjoint.

Figure 2 shows the recording setup at standard PC at the m&i Fachklinik Herzogenaurach. The audio data are collected using lapel microphones. One is attached to the clothes of the therapist (on the right) and one is attached to the patient’s clothes (on the left). This procedure allows segmenting the speech of both easily. All analyses are performed on the patient’s audio data only.

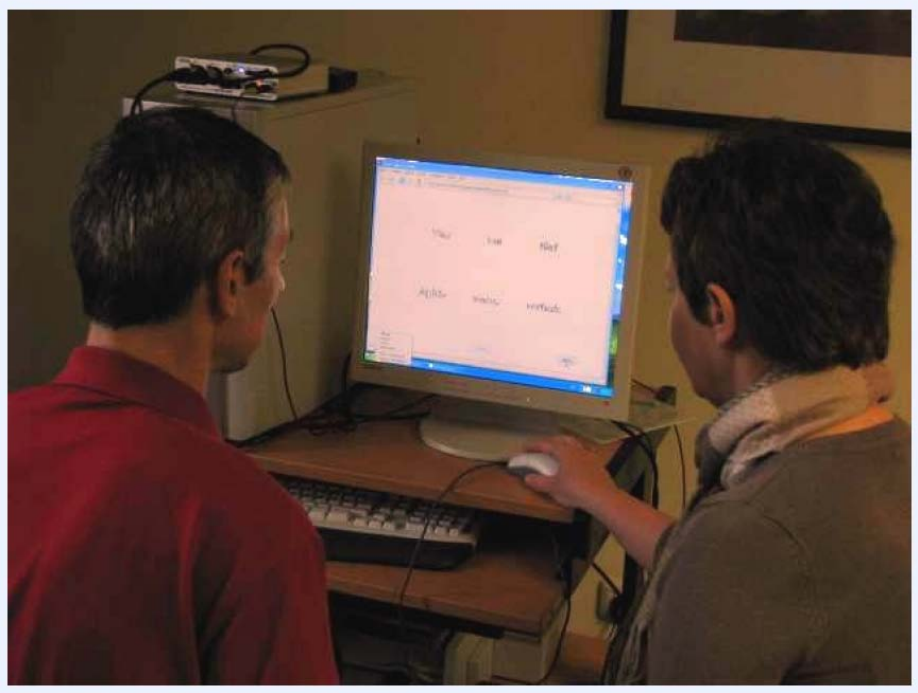

Figure 2: Recording setup at the m&i Fachklinik Herzogenaurach

4. Results

There was a correlation of r = -0.89 between the subjective assessment of the intelligibility and the word recognition rate of the speech recognition system ($p < 0.001$). The higher the recognition rate, the smaller the human score, because “1” means “very good”. The evaluations of the criterion “prosody” correlated with the ratio of the length of voiced and voiceless segments with r = 0.82 ($p < 0.01$). “Roughness” correlated with r = 0.81 with the average number of voiceless segments ($p < 0.01$). The correlation with “jitter” was only r = 0.66.

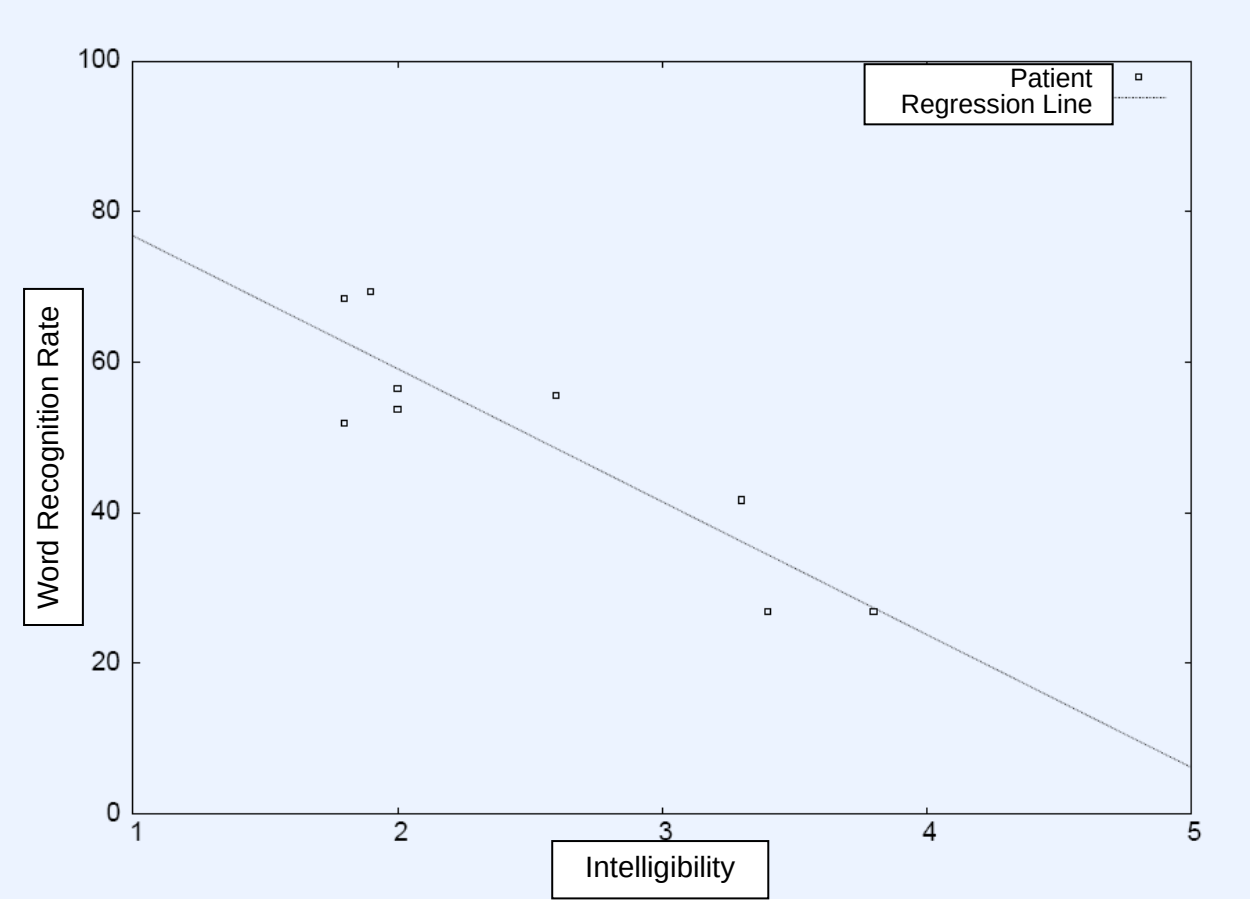

Figure 3: The correlation between the human Experts and the automatic system is high (r =−0, 89) and significant (p < 0, 001).

5. Discussion

Even in patients with dysarthria, a high and significant correlation between the word recognition rate of a speech recognition system and the perceptual assessment of the intelligibility is achieved. This is in line with previous studies [2].

Also on the criteria “prosody” and “roughness”, high and significant correlations between human evaluation and automatic prosodic features are found. The relationship of “prosody” and the ratio of the length of voiced and voiceless segments can be explained by the fact that both features are related to accentuation in speech. The correlation between “roughness” and the average number of voiceless segments is also plausible: A high roughness may disturb the automatic fundamental frequency extraction algorithm. This, in turn, causes an erroneous classification of voiced signals as voiceless. Hence, a long voiced segment may be divided into several short voiced and voiceless segments. Eventually, this increases the number of voiced and voiceless segments. This hypothesis is supported by the observation that the number of voiced segments also correlates at r = 0.80 ($p < 0.01$) with “roughness”.

The results as indicated by this first preliminary study indicate that the assessment of dysarthric speech is possible. If these results can be confirmed on a larger group of patients, the presented methods may allow

for objective and inter-independent analysis of the speech pathologies of dysarthric speakers. Hence, it will be possible to compare different therapy methods and the progress of a patient during therapy may be measured.

6. Summary

In this study, the automatic evaluation of dysarthric speech is investigated for the first time. In a small group of patients high and significant correlations between perceptual ratings of an expert panel and the evaluation by an automated system were achieved.

7. References

[1] S. Paal, U. Reulbach, K. Strobel-Schwarthoff, E. Nkenke, und M. Schuster. Beurteilung von Sprechauffälligkeiten bei Kindern mit Lippen-Kiefer-Gaumen-Spaltbildungen. J Orofac Orthop, 66(4):270–278, 2005.

[2] A. Maier, E. Nöth, A. Batliner, E. Nkenke, und M. Schuster, “Fully Automatic Assessment of Speech of Children with Cleft Lip and Palate,” Informatica, vol. 30, no. 4, pp. 477–482, 2006.

[3] G. Stemmer. Modeling Variability in Speech Recognition. Logos Verlag, Berlin, 2005.

[4] K. Pearson. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. Philosophical Transactions of the Royal Society of London, 187:253–318. 1896.