Automatic Evaluation of Pathologic Speech – from Research to Routine Clinical Use

Elmar Nöth[1], Andreas Maier[1,2], Tino Haderlein[1,2], Korbinian Riedhammer[1], Frank Rosanowski[2], and Maria Schuster[2]

[1] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5)
Martensstraße 3, 91058 Erlangen, Germany
noeth@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de
[2] Universität Erlangen-Nürnberg, Abteilung für Phoniatrie und Pädaudiologie
Bohlenplatz 21, 91054 Erlangen, Germany

Abstract. Previously we have shown that ASR technology can be used to objectively evaluate pathologic speech. Here we report on progress for routine clinical use: 1) We introduce an easy-to-use recording and evaluation environment. 2) We confirm our previous results for a larger group of patients. 3) We show that telephone speech can be analyzed with the same methods with only a small loss of agreement with human experts. 4) We show that prosodic information leads to more robust results. 5) We show that text reference instead of transliteration can be used for evaluation. Using word accuracy of a speech recognizer and prosodic features as features for SVM regression, we achieve a correlation of .90 between the automatic analysis and human experts.

1 Introduction

In speech therapy, objective evaluation of voice and speech quality is necessary for at least 1) patient assessment, 2) therapy control, 3) evaluation of different therapy methods using groups of patients, and 4) preventive screening. Normally, a group of experts rates some aspect of a patient's utterance like intelligibility, nasality, or harshness. This property is typically rated on a five to seven point Likert scale [1], e.g. from 1 = "very high" to 5 = "very low". The average or median of the ratings is then considered as an "objective" rating of this aspect of the patient's voice or speech. However, except for research projects, such a procedure is not done for financial, cost-cutting reasons. Thus, the patient is often evaluated by just one expert, sometimes only in a very crude way, e.g. the expert only distinguishes between "changed" and "unchanged intelligibility". With significant inter- and intra-rater variability, there normally is no objective evaluation of a patient's voice and speech available. Therefore, there is a strong need for an easy to apply, cost-effective, instrumental, and objective evaluation method.

In two research studies [2, 3] we showed that for two groups of patients such a method is available, using automatic speech recognition technology: For a group of children with "Cleft Lip and Palate" (CLP) we recorded names of objects shown on pictograms and for a group of patients with tracheoesophageal (TE) substitute voice (after removal of the larynx due to cancer) we recorded a

published in: Text, Speech and Dialogue (TSD 2007), Proceedings, LNAI 4629, Springer, Berlin, Heidelberg, 2007, pp. 294-301

phonetically rich read text. The recordings were transliterated and rated by a group of speech experts according to different aspects like intelligibility, nasality, and match of breath/sense units on a five-point Likert scale. The average of the intelligibility ratings was compared to the word accuracy (WA) of an automatic speech recognizer (ASR) which was calculated w.r.t. the transliteration. The correlation between these two ratings was .84 [3, 4] for TE and .9 [2] for CLP speech. When we projected the WA to the Likert scale and considered the ASR as an additional rater, the inter-rater agreement to the human raters was in the same range as the inter-rater agreement between the human raters. We can thus conclude that the WA can be used as an objective instrumental evaluation method.

In this paper we want to report on several steps that bring us closer to using ASR in everyday clinical use. In detail we will restrict ourselves to TE patients and will address the following topics which are all important steps towards a routine use of our evaluation methods:

1. Can we create an easy-to-use and easily available interface to our analysis environment?
2. Do our results hold for a larger, more representative group of patients?
3. A very important communication situation for the patient is the communication over the telephone, where other information channels are missing. Can we evaluate speech via this reduced information channel?
4. It is well known that prosody is an important aspect of speech perception. Can prosodic features improve our evaluation results?
5. To show the agreement between human experts and ASR, we carefully transliterated the utterances as a reference for WA. This would not be done outside of a research study. How well does the ASR rating agree to the human rating, when it is evaluated w.r.t. the reference text rather than the transliteration?

The rest of this paper is organized as follows: In Section 2 we give a short characteristic of TE voice and of our database. In Section 3 we introduce our recognition system and recording environment (topic 1). In Section 4 we try to answer the topics 2-5 named above. We conclude with a discussion and summary.

2 TE Voice and used Database

The TE substitute voice is currently state-of-the-art treatment to restore the ability to speak after laryngectomy [5]: A silicone one-way valve is placed into a shunt between the trachea and the esophagus which on the one hand prevents aspiration and on the other hand deviates the air stream into the upper esophagus during expiration. The upper esophagus, the pharyngo-esophageal (PE) segment, serves as a sound generator. Tissue vibrations of the PE segment modulate the streaming air and generate the primary substitute voice signal which is then further modulated in the same way as normal speech. In comparison to normal voices the quality of substitute voices is low, e.g. the change of pitch and volume is limited and inter-cycle frequency perturbations result in a hoarse voice [6]. Acoustic studies of TE voices can be found for instance in [7, 8].

41 laryngectomees ($\mu = 62.0 \pm 7.7$ years old, 2 female and 39 male) with TE substitute voice read the German version of the text "The North Wind and

the Sun", a fable from Aesop. It is a phonetically rich text with 108 words (71 disjoint) which is often used in speech therapy in German speaking countries. The speech samples were recorded with a close-talking microphone with 16 kHz and 16 bit.

To determine the loss of information due to the telephone channel, we played back the close-talking recordings using a standard PC and loudspeaker in a quiet office environment and placed a telephone headset in front of the loudspeaker, i.e. we created a telephone quality (8 kHz a-law) version of the database. Due to the multiple AD/DA conversions and the different frequency characteristics of the loudspeaker and the microphones we expect the recognition rates to be a lower bound for the recognition rates for real telephone calls.

3 The Automatic Speech Analysis System

For the objective measurement of the intelligibility of pathologic speech, we use a hidden Markov model (HMM) based ASR system. It is a state-of-the-art word recognition system developed at the Chair of Pattern Recognition (Lehrstuhl für Mustererkennung) of the University of Erlangen-Nuremberg. In this study, the latest version as described in detail in [9] was used. A commercial version of this recognizer is used in high-end telephone-based conversational dialogue systems by *Sympalog* (www.sympalog.com), a spin-off company of the Chair of Pattern Recognition. As features we use 11 Mel-Frequency Cepstrum Coefficients and the energy of the signal for a 16 ms analysis frame (10 ms shift). Additionally 12 delta coefficients are computed over a context of 2 time frames to the left and the right side (56 ms in total). The recognition is performed with semi-continuous HMMs. The codebook contains 500 full covariance Gaussian densities which are shared by all HMM states. The elementary recognition units are polyphones [10], a generalization of triphones.

The output of the word recognition module is used by our prosody module to calculate word-based prosodic features. Thus, the time-alignment of the recognizer and the information about the underlying phoneme classes (like *long vowel*) can be used by the prosody module. For each word we extract 22 prosodic features over intervals of different sizes, i.e. the current word or the current word and the previous word. These features model F0, energy and duration, e.g. maximum of the F0 in the word pair "current word and previous word". In addition, 15 global prosodic features for the whole utterance are calculated, e.g. standard deviation of jitter and shimmer. In order to evaluate the pathologic speech, we calculate the average, the maximum, the minimum, and the variance of the 37 turn- and word-based features for the whole text to be read. Thus we get 148 features for the whole text. A detailed description of the features is beyond the scope of this paper. We will restrict ourselves to explaining in Section 4 those features which proved to be most relevant for our task. A detailed discussion of our prosodic features can be found in [11, 12].

3.1 Recognizer Training

The basic training set for our recognizers are dialogues from the VERBMOBIL project [13]. The topic of the recordings is appointment scheduling. The data

were recorded with a close-talking microphone with 16 kHz and 16 bit. The speakers were from all over Germany and thus covered most dialect regions. However, they were asked to speak standard German. About 80% of the 578 training speakers (304 male, 274 female) were between 20 and 29 years old, less than 10% were over 40. This is important in view of the test data, because the fact that the average age of our test speakers is more than 60 years may influence the recognition results. A subset of the German VERBMOBIL data (11,714 utterances, 257,810 words, 27 hours of speech) was used for the training set and 48 utterances (1042 words) for the validation set (the training and validation corpus was the same as in [9]).

In order to get a telephone speech recognizer, we downsampled the training set to telephone quality. We reduced the sampling rate to 8 kHz and applied a low-pass filter with a cutoff frequency of 3400 Hz to simulate telephone quality. Thus, we used "the same" training data for the close-talking and telephone recognizer. A loss in evaluation quality will therefore mainly be caused by the different channels, not by different amounts of training data.

In [4], we showed for a corpus of 18 TE speakers that a monophone-based recognizer for close-talking signals produced slightly better agreement with speech experts' intelligibility ratings than a polyphone-based recognizer. We wanted to verify all the results for the larger corpus of 41 TE speakers. Therefore we created four different recognizers: For the 16 kHz and the 8 kHz training data, we created a polyphone-based and a monophone-based recognizer (rows "16/m", "8/m", "16/p", "8/p" in Table 2). After the training, the vocabulary was reduced to the words occurring in the German version of the "The North Wind and the Sun".

3.2 Recording Environment

For routine use of our evaluation system, it must be easily and cheaply available from any phoniatric examination room. We created PEAKS (**P**rogram for **E**valuation and **A**nalysis of all **K**inds of **S**peech disorders), a client/server recording environment. The system can be accessed from any PC with internet access, a browser, a sound card, and Java Runtime Environment (JRE) 1.5.0.6. The texts to be read and pictograms to be named are displayed in the browser. The patient's utterances are recorded by the client and transferred to the server. The ASR system analyzes the data and sends the evaluation results back to the client. The recordings are stored in an SQL database. A secure connection is used for all data transfer. A registered physician can group his patients according to disorder, create new patient entries, create new recordings, analyze patients and groups of patients. The physician has only access to his patients but physicians can share groups of patients. For the telephone data, the patient gets a handout from his physician with a unique ID and the text to be read. The server can be accessed from the public telephone system. PEAKS is used by physicians from 3 clinics of our university, collecting data from patients with CLP, TE voice, epithelium cancer in the oral cavity, and partial laryngectomy. More information can be found at
http://www5.informatik.uni-erlangen.de/Research/Projects/Peaks.

4 Experimental Results

4.1 Subjective Evaluation

A group of 5 voice professionals subjectively estimated the intelligibility of the 41 patients while listening to a play-back of the close-talking recordings. A five-point Likert scale was applied to rate the intelligibility of each recording. In this manner an averaged mark – expressed as a floating point value – for each patient could be calculated. We assigned this mark also to the telephone recordings.

To judge the agreement between the different raters we calculated correlation coefficients and the weighted multi-rater κ [14] for the "intelligibility" rating. The average correlation coefficient between a single rater and the average of the 4 other raters was .81, the weighted multi-rater κ for the 5 raters was .45. A κ value greater than .4 is said to show moderate agreement.

4.2 Automatic Evaluation

We applied the two close-talking recognizers and the two telephone speech recognizers to the accordant speech data and calculated the correlation between the WAs and the average of the experts' intelligibility rating. The WA was calculated w.r.t. the reading text and w.r.t. the transliteration. The κ values were calculated using the recognizer as a 6th rater. For this we mapped the WAs to the Likert scale, using the thresholds that are given in Table 1.

Mark	5	4	3	2	1	Mark	5	4	3	2	1
WA c/t	< 5	< 25	< 40	< 55	≥ 55	WA tel	< 5	< 15	< 25	< 45	≥ 45

Table 1. Thresholds for mapping the WA of the close-talking (c/t) and the telephone (tel) ASR systems to marks on the Likert scale for rating the intelligibility of the patients.

In a second step we applied a 10-fold cross-validation multi correlation/regression analysis [15] to determine the features with the best average rank among WA and the 148 prosodic features. These features are either global or the average features calculated for words or word pairs (see above and [11, 12]).

We used these features and the average expert rating for SVM regression [16, 17]. Rounding the SVM regression value to the next integer we again treated the automatic result as a 6th rater and calculated the multi-rater κ.

The multi correlation/regression analysis chose the following features (in descending order):

- WA always had the the best rank.
- The global F0 mean.
- The variance of the energy maximum.
- The maximum pause duration before a word.
- The mean of the F0 regression coefficient.

In this work, only the first two features were used due to the small size of the test set.

Table 2 shows the results for the 4 recognizers based on WA and WA in combination with the best prosodic feature (P). Note that r_{WA} is negative, since good speakers have low Likert values and high WAs, while r_P is positive since SVM regression tries to predict the average score of the human raters. Figure 1 shows the SVM regression values vs. the average experts' score as well as the regression line. The result of the 16/p recognizer and the text reference were used for the calculation of the WA.

reco	eval	μ_{WA}	r_{WA}	κ_{WA}	r_P	κ_P	reco	eval	μ_{WA}	r_{WA}	κ_{WA}	r_P	κ_P
16/m	trl	37.7	-.85	.44	—	—	16/p	trl	38.6	-.89	.47	.89	.48
8/m	trl	31.1	-.78	.38	—	—	8/p	trl	27.5	-.84	.40	—	—
16/m	ref	37.7	-.85	.44	—	—	16/p	ref	38.6	-.89	.46	.90	.47
8/m	ref	31.0	-.78	.38	—	—	8/p	ref	27.5	-.83	.41	—	—

Table 2. Evaluation results for the four different recognizers for the 41 patients. The WA is calculated w.r.t. the transliteration (trl) and text reference (ref), r is the correlation between the WA or the SVM regression and the average expert rating. For the description of the recognizers see Section 3.1

5 Discussion and Summary

In the following we want to discuss the topics addressed in Section 1.

1. The recording environment is highly accepted by our clinical colleagues. One major reason is that there is no installation cost, since practically all examination rooms already have a telephone and a PC with internet access. We are currently expanding the data collection to other German clinics.
2. The results reported on 18 patients in [3, 4] were mostly confirmed for the 41 patients. The best correlation (-.89) and κ values (.47) were slightly higher than for the 18 patients (-.84 and .43). For the larger corpus, the polyphone-based recognizers produced better and more consistent results than the monophone-based ones. Thus, our assumption that the monophone models are more robust towards the strongly distorted TE speech [4] seems not to hold.
3. The results for the telephone recognizers show that the loss of information due to the telephone channel are acceptable, e.g. from -.89 and .47 for "16/p" to -.84 and .40 for "8/p", respectively. Due to the loss of quality in telephone transmission, the multiple AD/DA conversions, and the different frequency characteristics of the loudspeaker and the microphones, the overall WA for the simulated telephone calls is reduced. Also, the training data of the speech recognizer for the 8 kHz was downsampled close-talking data and not real telephone data. We chose this way instead of using real telephone training data, since we wanted the telephone recognizer to be trained with the same training data as the recognizer for the close-talking data. Reducing the

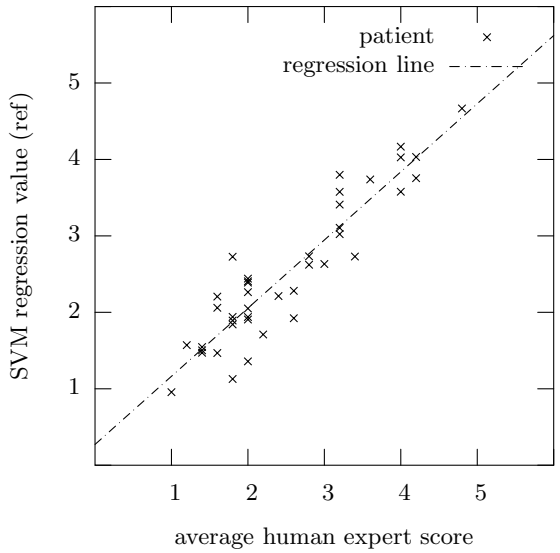

Fig. 1. SVM regression value for the 41 recordings in comparison to the average of the experts' intelligibility score.

acoustic mismatch of training and evaluation data might lower the loss of correlation.

4. Adding prosodic features to the evaluation vector increases the correlation to the human experts' scores (from .89 to .90) and makes the analysis more robust. We are currently porting the prosody module to telephone speech.
5. The results in Table 2 show that there is practically no difference between the results evaluated against the transliteration and against the reference text. Thus we can do without the cumbersome transliteration.

In conclusion we can say that our evaluation system provides an easy to apply, cost-effective, instrumental, and objective evaluation for TE speech. We are currently enhancing our analysis environment in order to provide a modular platform which can be easily expanded:

- From the medical point of view we can add new intelligibility tests to provide speech evaluation for a larger spectrum of speech disorders. The easy to use graphical user interface allows a fast evaluation of these tests.
- From the technical point of view we are able to plug in different ASR systems in order to provide more flexibility when realizing these new tests.
- Once a new intelligibility test is integrated and validated, it can immediately be used in clinical routine in all clinics participating. Thus PEAKS not only speeds up research studies but also helps to reduce the gap between research and practice.

Acknowledgments

This work was funded by the Deutsche Krebshilfe (German Cancer Aid) under grant 106266, the Deutsche Forschungsgemeinschaft (German Research Foundation) under grant SCHU2320/1-1, and the Johannes-und-Frieda-Marohn Stiftung. The responsibility for the content of this paper lies with the authors.

References

1. Likert, R.: A Technique for the Measurement of Attitudes. Archives of Psychology **140** (1932)
2. Schuster, M., Maier, A., Haderlein, T., Nkenke, E., Wohlleben, U., Rosanowski, F., Eysholdt, U., Nöth, E.: Evaluation of Speech Intelligibility for Children with Cleft Lip and Palate by Means of Automatic Speech Recognition. International Journal of Pediatric Otorhinolaryngology **70** (2006) 1741–1747
3. Schuster, M., Haderlein, T., Nöth, E., Lohscheller, J., Eysholdt, U., Rosanowski, F.: Intelligibility of Laryngectomees' Substitute Speech: Automatic Speech Recognition and Subjective Rating. European Archives of Oto-Rhino-Larngology and Head & Neck **263** (2006) 188–193
4. Schuster, M., Nöth, E., Haderlein, T., Steidl, S., Batliner, A., Rosanowski, F.: Can you Understand him? Let's Look at his Word Accuracy — Automatic Evaluation of Tracheoesophageal Speech. (Volume 1.) 61–64
5. Brown, D., Hilgers, F., Irish, J., Balm, A.: Postlaryngectomy Voice Rehabilitation: State of the Art at the Millennium. World J Surg **27**(7) (2003) 824–831
6. Schutte, H., Nieboer, G.: Aerodynamics of esophageal voice production with and without a Groningen voice prosthesis. Folia Phoniatrica et Logopaedia **54** (2002) 8–18
7. Robbins, J., Fisher, H., Blom, E., Singer, M.: A Comparative Acoustic Study of Normal, Esophageal, and Tracheoesophageal Speech Production. Journal of Speech and Hearing Disorders **49** (1984) 202–210
8. Bellandese, M., Lerman, J., Gilbert, H.: An Acoustic Analysis of Excellent Female Esophageal, Tracheoesophageal, and Laryngeal Speakers. Journal of Speech, Language, and Hearing Research **44** (2001) 1315–1320
9. Stemmer, G.: Modeling Variability in Speech Recognition. Volume 19 of Studien zur Mustererkennung. Logos Verlag, Berlin (2005)
10. Schukat-Talamazzini, E., Niemann, H., Eckert, W., Kuhn, T., Rieck, S.: Automatic Speech Recognition without Phonemes. In: Proc. European Conf. on Speech Communication and Technology. Volume 1., Berlin (1993) 111–114
11. Batliner, A., Buckow, A., Niemann, H., Nöth, E., Warnke, V.: The Prosody Module. [13] 106–121
12. Haderlein, T., Nöth, E., Schuster, M., Eysholdt, U., Rosanowski, F.: Evaluation of Tracheoesophageal Substitute Voices Using Prosodic Features. In: Proc. of 3rd International Conference on Speech Prosody, Dresden (2006) 701–704
13. Wahlster, W., ed.: Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin (2000)
14. Davies, M., Fleiss, J.: Measuring agreement for multinomial data. Biometrics **38**(4) (1982) 1047–1051
15. Cohen, J., Cohen, P.: Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale, New Jersey (1983)
16. Platt, J.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Schölkopf, B., Burges, C., Smola, A., eds.: Advances in Kernel Methods – Support Vector Learning. MIT Press (1999) 185–208
17. Smola, A., Schölkopf, B.: A Tutorial on Support Vector Regression. In: NeuroCOLT2 Technical Report Series. (1998) NC2-TR-1998-030.