TOWARDS ROBUST AUTOMATIC EVALUATION OF PATHOLOGIC TELEPHONE SPEECH

K. Riedhammer[1], *G. Stemmer*[2], *T. Haderlein*[1,3], *M. Schuster*[3], *F. Rosanowski*[3], *E. Nöth*[1], *A. Maier*[1,3]

[1]Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg
Martensstraße 3, 91058 Erlangen, GERMANY
maier@informatik.uni-erlangen.de

[2]Siemens AG, Corporate Technology, CT IC5
Otto-Hahn-Ring 6, 81730 München, GERMANY
georg.stemmer@siemens.com

[3]Abteilung für Phoniatrie und Pädaudiologie, Universität Erlangen-Nürnberg
Bohlenplatz 21, 91054 Erlangen, GERMANY

ABSTRACT

For many aspects of speech therapy an objective evaluation of the intelligibility of a patient's speech is needed. We investigate the evaluation of the intelligibility of speech by means of automatic speech recognition. Previous studies have shown that measures like word accuracy are consistent with human experts' ratings. To ease the patient's burden, it is highly desirable to conduct the assessment via phone. However, the telephone channel influences the quality of the speech signal which negatively affects the results. To reduce inaccuracies, we propose a combination of two speech recognizers. Experiments on two sets of pathological speech show that the combination results in consistent improvements in the correlation between the automatic evaluation and the ratings by human experts. Furthermore, the approach leads to reductions of 10% and 25% of the maximum error of the intelligibility measure.

Index Terms— Biomedical acoustics, Speech intelligibility, Speech processing, Acoustic applications

1. INTRODUCTION

In speech therapy objective evaluation of voice and speech quality is required for patient assessment, therapy control, comparative evaluation of different therapy methods, and preventive screening. Conventionally a group of experts rates a specific aspect such as intelligibility or nasality of a patient's utterance. The ratings are usually on a five- to seven-point Likert scale [1], e. g. reaching from "very high" to "very low". The average, median, or majority of the ratings of all experts is then considered to be an "objective" rating of the patient's voice or speech. Unfortunately, except for research projects, such a procedure is frequently refrained from due to the high effort in time and finance: the patient's voice is often evaluated by just a single expert and sometimes even in a very inaccurate way (e. g. "altered" or "not altered"). Therefore there is a strong need for an instrumental and objective evaluation method that is easy to apply and cost-effective.

Previous studies [2, 3, 4] have demonstrated the effectiveness of a new objective evaluation method: the patient reads a phonetically rich text and an automatic speech recognition (*ASR*) system is applied to the recorded speech data; word accuracy *WA* or word recognition rate *WR* are computed for the recognizer's output w. r. t. the reference text. It has been shown for three types of voice and speech disorders (tracheoesophageal, *TE*, speech [2], speech of children with cleft lip and palate, *CLP* [3], and speech of patients treated for cancer of the oral cavity, *OC* [4]) that these automatically computed measures have high consistency with the experts' evaluations, i. e. the amount of errors made by a speech recognizer is highly correlated to the experts' rating of the intelligibility of a patient's voice.

The goal of the work described in this paper is to develop the approach further to allow for the automatic evaluation of a patient over a telephone connection. Examinations via phone can ease the patients' burden in case of limited mobility or far distances as there is no need to travel to a specialist. Due to the low cost, the technique is affordable even for practitioners with a small budget. Additionally, as the telephone is important in our daily life, it could be an aspect of speech therapy to evaluate and to improve in particular the patient's intelligibility in telephone conversations. Previous work on tracheoesophageal speech [5] showed that evaluation over telephone is feasible. However, due to the fact that telephone transmission reduces the signal quality and bandwidth, it was observed that the correlation between the automatically generated measures and the experts' rating is lower than for recordings made with close-talk microphones.

Up to now, experiments were performed using a single ASR system. Having in mind that the human objective rating is a combination of several experts, the following questions arise: Can a combination of several ASR systems outperform a single ASR system in means of agreement with the human experts group? Additionally, can the loss of signal quality due to the telephone transmission be compensated in the same way? To answer these questions, we investigate different approaches to combine two different recognizers optimally, either using the generated word hypotheses or based on the generated measurements. We aim to maximize the correlation between the human experts and the automatic evaluation methods. Another important aspect that has not yet been considered in our previous work is the maximum error of the generated measurements, i. e. the maximum difference between an automatically generated rating and the experts' label for the same speaker which should be as small as possible. Experiments are conducted on two databases collected from patients with tracheoesophageal substitute voice and patients

with cancer of the oral cavity which have already been investigated in previous works.

The paper is organized as follows: Section 2 gives an overview over the data sets used throughout this paper. In Section 3, the employed speech recognition systems are introduced. Section 4 describes how the two ASR systems are combined, and Section 5 investigates the experimental results. We conclude with a discussion and outlook on future work in Section 6.

2. SPEECH DATA AND SUBJECTIVE EVALUATION

The experiments described in this paper are conducted on two data sets recorded from speakers with two different types of pathological voices. Both data sets are recordings of 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 commonly used in German speech therapy.

2.1. The *OC* Data Set

The first data set consists of 46 patients treated for cancer of the oral cavity ($\mu = 60 \pm 10$ years old, 13 female and 33 male). After surgical treatment, they suffer from various functional restrictions such as speech disorders with a high individual variability. The patients' utterances were acquired using PEAKS [4] with a close-talk microphone at a sampling rate of 16 kHz and 16 bit quantization. A detailed description of how the speech data were recorded can be found in [4]. The text was split into ten turns at major syntactic boundaries in order to be able to display the text in large letters well readable for elderly people without disturbing the reading flow. The software automatically segments the audio data according to these boundaries. In order to get an intelligibility score for the patients, 4 experts rated the recordings on a 5 point Likert scale reaching from 1 ="very good" to 5 = "very bad". The final intelligibility score for each patient is obtained by averaging the marks of all turns and experts. In order to get a telephone quality version of this data set, the recordings were downsampled to 8 kHz and low-pass filtered at 4 kHz. As the human intelligibility ratings are expected to be independent of telephone transmission or to be at least consistently lower for all patients, their ratings were just copied. Of course, this procedure does not guarantee real telephone quality, but can be taken as an upper border of telephone signal quality. In the following, this data set is referenced as *OC*. Throughout the paper we will refer to the original 16 kHz data as *OC-orig*.

2.2. The *TE* Data Set

The second data set consists of 41 patients with tracheoesophageal (*TE*) substitute voices ($\mu = 62 \pm 8$ years old, 2 female and 39 male). The TE substitute voice is a treatment to restore the ability to speak after laryngectomy, i. e. the larynx (including vocal folds) is removed, the upper end of the trachea is closed, and an artificial exit (*tracheostoma*) is created to allow breathing. A silicone shunt valve is used to connect the trachea and the esophagus which allows, once the tracheostoma is occluded, to deviate the air flow into the pharyngo-esophageal segment. There, tissue vibrations generate the primary 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. Refer to [5] for a detailed description of the acoustic properties of the data set. The recordings were acquired using a sampling rate of 16 kHz and 16 bit quantization. Afterwards, they were presented to a group of 5 experts who judged the intelligibility using the same Likert scale from above. These ratings were combined to an average rating per patient. The audio data were played back into a telephone in a quiet office room and recorded again at the other end of the line in order to get a telephone quality version of the data (8 kHz, 16 bit). Due to the same reasons as given above, the intelligibility scores are taken from the unfiltered data. In contrast to the *OC* data set, the signal quality of this corpus is considered as a lower border of telephone signal quality, as the utterances were played back through a loudspeaker in addition to the real telephone transmission. In the following, this data is referred to as *TE* while the original 16 kHz corpus is referred to as *TE-orig*.

3. AUTOMATIC SPEECH RECOGNITION SYSTEMS

Two different baseline speech recognizers have been employed for the experiments described in this paper: the speech recognizer from the *Universität Erlangen-Nürnberg* (*ER*) and a research system from *Siemens AG* (*SIE*).

3.1. Universität Erlangen-Nürnberg

The *ER* automatic speech recognizer has been developed at the *Lehrstuhl für Mustererkennung* of the *Universität Erlangen-Nürnberg*. A commercial version of the system is used in high-end telephone-based conversational dialogue systems by the spin-off company *Sympalog* (www.sympalog.com). The feature set consists of Mel-frequency cepstral coefficients and the energy together with the corresponding deltas. Semi-continuous Hidden Markov Models (HMM) are employed as acoustic models with a single codebook of 500 full covariance Gaussian densities. The elementary recognition units are polyphones, a generalization of triphones. In order to focus on acoustic properties of the speakers' voices, only a unigram language model is used. The *ER* speech recognizer has been trained on dialogues from the VERBMOBIL project. The topic of the recordings is appointment scheduling. The data were recorded with a close-talk microphone with 16 kHz and 16 bit. The speakers were from all over Germany and thus covered most regions of dialect and had no voice or speech impairment. A subset of the German VERBMOBIL data (11,714 utterances, 257,810 words, 25 hours of speech) was used for training and 48 utterances (1042 words) for validation. A detailed description of the recognizer and the training and test set can be found in [6]. 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 4 kHz to simulate telephone transmission. Thus, we used "the same" training data for the close-talk and telephone recognizer as in [4, 5]. A loss in evaluation quality will therefore mainly be caused by channel differences rather than by different amounts of training data.

3.2. Siemens AG

The *SIE* ASR system was developed at the *Professional Speech Processing (CT IC5)* group of *Siemens AG, Corporate Technology*. The front-end of the *SIE* recognizer combines Mel-frequency cepstral coefficients and energy into a feature vector which is then concatenated with several neighboring frames and reduced to a dimension of 24 with a linear feature transformation. The acoustic models are state-tied, gender-independent continuous triphone HMMs. A phonetic decision tree is used for tying states and defining the

context-dependent allophones. The baseline system has about 1500 tied states and about 14000 Gaussians. The Gaussian distributions differ only by the mean vectors, there is just one global variance parameter. For training of the recognizer, various corpora of German telephone speech have been combined which sum up to about 160 hours of speech of about 2000 speakers. For the same reasons as described above during recognition a unigram language model is employed.

4. COMBINING RECOGNIZERS

To improve the automatic intelligibility evaluation we propose to utilize a suitable combination two different speech recognizers. This is motivated by the observation that independent speech recognition systems make different errors and that their combination may thus reduce inaccuracies caused by recognition errors. We investigate two approaches: the first is based on the direct combination of the recognizer outputs, i. e. of the generated word hypotheses, while the second method combines the recognition rates computed for both systems.

4.1. Combining Recognizer Outputs

A well-known algorithm for the combination of the output of multiple ASR systems is the ROVER technique [7]. Usually, ROVER computes the most likely word chain using confidence measures for the single word hypotheses. In our case, we feed the ROVER algorithm with three word hypotheses: the reference text and the two recognizer outputs, all with the same confidence values. This leads to the effect that if at least one recognizer output contains the correct word, it gets a majority voting in combination with the reference text. Hence, a word chain is generated out of the two recognizer outputs which matches the reference text best. For this output, the word accuracy

$$\mathrm{WA} = \left(1 - \frac{D+S+I}{R}\right) \cdot 100\%$$

and the word recognition rate

$$\mathrm{WR} = (D+S)/R \cdot 100\%$$

can be computed as if it were an independent recognizer output. Note that D is the number of deleted, S the number of substituted, I the number of inserted and R the number of reference words.

4.2. Combining Recognition Rates

In contrast to combining recognizer outputs, recognition rates like WA and WR can also be combined. The combination might not be that vulnerable to outliers as for this both recognizers would have to fail. Hence, the WA and WR computed on the recognizer outputs and the ROVER word chain are used as a kind of meta features for each patient and recording. As WA, WR and the human average ratings are continuous numbers, the problem can be formulated as a regression problem. A very simple form of regression is the linear regression (*LR*). It is defined as

$$y = f(\boldsymbol{x}) = b + \sum_i w_i x_i = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b \tag{1}$$

where $\boldsymbol{x}$ denotes the i-dimensional input vector (in this case made up of WAs and WRs of different word chains), $\boldsymbol{w}$ denotes the weight vector and y denotes the target regression value (in this case the average human intelligibility rating). Given a set of N training data samples $(\boldsymbol{x}_j, y_j)$, the optimization problem is usually to find a weight vector $\boldsymbol{w}$ that minimizes the sum of squared errors

$$\varepsilon = \sum_{j=1}^{N} (y_j - f(\boldsymbol{x}_j)) \quad . \tag{2}$$

Once $\boldsymbol{w}$ is estimated and given an input vector $\boldsymbol{x}$, the regression model (1) can be used to predict y. The LR used for the experiments of this work is provided by the data analysis tool WEKA [8]. The implementation is based on [9].

LR can be sensitive to outliers as every training sample equally contributes to the solution of the optimization problem. A more robust type of regression is support vector regression (*SVR*). For the basic linear SVR, the objective formula of the regression is the same as (1). The difference is in the optimization problem. The idea is to find a solution which permits small outliers i. e., prediction errors smaller than ε are tolerated. Additionally, to handle errors greater than ε, *slack variables* ξ_j, ξ_j^* are introduced to weight these outliers. The optimization problem can be formulated as

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{j=1}^{N}(\xi_j + \xi_j^*) \\ &\text{subject to} && \begin{cases} y_j - \langle \boldsymbol{w}, \boldsymbol{x}_j \rangle - b \le \varepsilon + \xi_j \\ \langle \boldsymbol{w}, \boldsymbol{x}_j \rangle + b - y_j \le \varepsilon + \xi_j^* \\ \xi_j, \xi_j^* \ge 0 \end{cases} \end{aligned} \tag{3}$$

The constant $C > 0$ determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated. This corresponds with the so called ε-insensitive loss function $|\xi|_\varepsilon$ described by

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{if } |\xi| \le \varepsilon \\ |\xi| - \varepsilon & \text{otherwise.} \end{cases} \tag{4}$$

The solution of this optimization problem leads to the result that the weight vector $\boldsymbol{w}$ is made up of a weighted combination of the training samples. Hence, every training pattern can individually contribute to the solution of the optimization. For more details, please refer to [10] which is also the basis for the implementation provided by WEKA.

We estimate a regression model using a training set labeled by human experts and use this model to predict new intelligibility ratings based on the automatically extracted measures provided by the ASR systems. Although there are only 6 meta features per patient and recording (WA and WR of 3 word chains), a feature selection is reasonable as not all features might be useful for the regression. Therefore, a data driven feature selection provided by WEKA is applied: using subsets of the data and a best-first search, the most valuable features are determined. For more details, please refer to [11].

The algorithms above require feature selection and training using labeled data. However, the data sets presented in Section 2 consist of only 41 and 46 individuals. As this is way to less for a common train/test split using 80% / 20%, a leave-one-out (*LOO*) approach is used to get the most out of the small data set size. Note that LOO is only used for the SVR experiments. The correlations with the LR results are computed with equal train and test sets as this is the normal procedure to measure correlation between two variables.

5. EXPERIMENTAL RESULTS

We present results for the data sets *OC* and *TE* introduced in Sec. 2 using the recognizers *ER* and *SIE* described in Sec. 3. The goal of our work is to be able to perform assessments of pathological speech

	ER	*SIE*	*ROV*
min(WA)	0	-25.9	49.1
max(WA)	84.3	84.7	97.2
mean(WA)	42.7	42.8	84.1
min(WR)	15.7	11.1	78.7
max(WR)	84.3	88.9	100
mean(WR)	48.1	53.0	91.5

Table 1. Detailed recognition results WA and WR in percent for the *ER* and *SIE* recognizers and the ROVER combination *ROV* using the *OC* data set.

	Train = Test		LOO	
	r(LR)	ρ(LR)	r(SVR)	ρ(SVR)
SIE	0.89	0.88	0.87	0.86
ER	0.89	0.88	0.88	0.86
SIE+ER	**0.93**	**0.90**	**0.92**	**0.90**
ER/16 kHz	0.92	0.91	0.92	0.90

Table 2. Correlation of the WA of the *ER* recognizer, the WR of the *SIE* recognizer, and the combination of the two values to the average human intelligibility rating using the *OC* data set. The ER/16 kHz results are based on the WR and the *OC-orig* data set. Note that the SVR results computed using LOO are nearly as good as the results using LR and a train = test scenario. (r – Pearson correlation coefficient, ρ – Spearman rank correlation coefficient)

over the telephone without a significant degradation in the quality of the ratings. Therefore we compare the individual systems and their combinations on telephone speech with the performance that has been achieved by a single recognizer on 16 kHz data recorded with close-talk microphones.

5.1. Results for the *OC* Data Set

For the *OC* data set, the detailed recognition results are shown in Table 1. The recognition rates with ROVER are very high (WA mean: 84.1%) which is intuitive as it compensates the individual errors of the recognizers using the reference text. To determine the best features to combine among these measures, data driven feature selection was used. In a LOO experiment, the feature selection provided by WEKA always chose at least the WA of the *ER* and the WR of the *SIE* recognizer. The ROVER values did not show a good correlation to the human average score and were therefore not selected by the feature selection algorithm. Hence, only WA of the *ER* and the WR of the *SIE* recognizer are used in the following.

Using the SVR and a LOO evaluation, the WA of the *ER* and the WR of the *SIE* recognizer independently show a Pearson correlation [12] of $r = 0.88$ (Spearman correlation [13]: $\rho = 0.86$) and 0.87 (0.86) to the human average intelligibility score (see Table 2). The combination of both results increases the correlation to 0.92 (0.90) which is in the same range of the 16 kHz experiment in [4]. Hence, the information lost due to the simulated telephone channel could be recovered. Additionally, the average absolute error was reduced from 0.37 (*ER*) / 0.43 (*SIE*) to 0.31 (see Table 3). The maximum absolute error was reduced to 0.94 which is higher as the maximum absolute error of the *SIE* recognizer, but still in a good range.

	ER	SIE	SIE+ER	ER/16 kHz
	OC			*OC-orig*
average error	0.37	0.43	**0.31**	0.31
minimum error	0.01	0.06	**0.03**	0
maximum error	1.25	0.86	**0.94**	0.92
	TE			*TE-orig*
average error	0.49	0.42	**0.42**	0.38
minimum error	0.01	0	**0.04**	0.02
maximum error	1.21	1.27	**1.09**	1.05

Table 3. Average absolute, minimum and maximum error for the single recognizers and the combination using SVR and LOO. The results for the *OC* data set are computed using WA of the *ER* and WR of the *SIE* recognizer. The results for the *TE* data set are computed using WA of the *ER* and *SIE* recognizer. The ER/16 kHz results are computed using the WA for the *TE-orig* and WR for the *OC-orig* data set.

	ER	*SIE*	*ROV*
min(WA)	-15.7	-29.6	48.1
max(WA)	54.6	75.9	92.6
mean(WA)	27.6	36.3	79.2
min(WR)	13.0	21.3	82.4
max(WR)	63.0	82.4	98.1
mean(WR)	38.7	52.2	91.9

Table 4. Detailed recognition results WA and WR in percent for the *ER* and *SIE* recognizers and the ROVER combination *ROV* using the *TE* data set.

5.2. Results for the *TE* Data Set

For the *TE* data set, the detailed recognition results are shown in Table 4. In a LOO experiment, the feature selection always chose at least the WA of both recognizers. Like with the *OC* data set, the ROVER recognition results are very high but show nearly no correlation to the average human rating. Hence, only the WA of both recognizers is considered in the following.

Using SVR and a LOO evaluation, the WA of both recognizers independently yield a correlation of 0.82 (0.82) and 0.79 (0.81) to the average human intelligibility score (see Table 5). The combination of both recognizers increases the correlation to 0.84 (0.84). The WA computed on the 16 kHz data and the technically equal 16 kHz recognizer showed a correlation of 0.89 and is considered as an upper border [5]. Hence, the correlation of the combination of the two recognizers approaches this upper border. Additionally, the maximum absolute error was reduced from 1.21 (*ER*) / 1.27 (*SIE*) to 1.09, and the average absolute error was reduced to 0.42 (see Table 3).

6. DISCUSSION AND FUTURE WORK

We show for two telephone speech corpora with different voice and speech pathologies that a combination of two independent automatic speech recognition systems outperforms a single ASR system when comparing recognition rates like WA or WR to human average intelligibility scores. Additionally, the average and maximum absolute errors could be reduced which is important when it comes to clinical use of this technique.

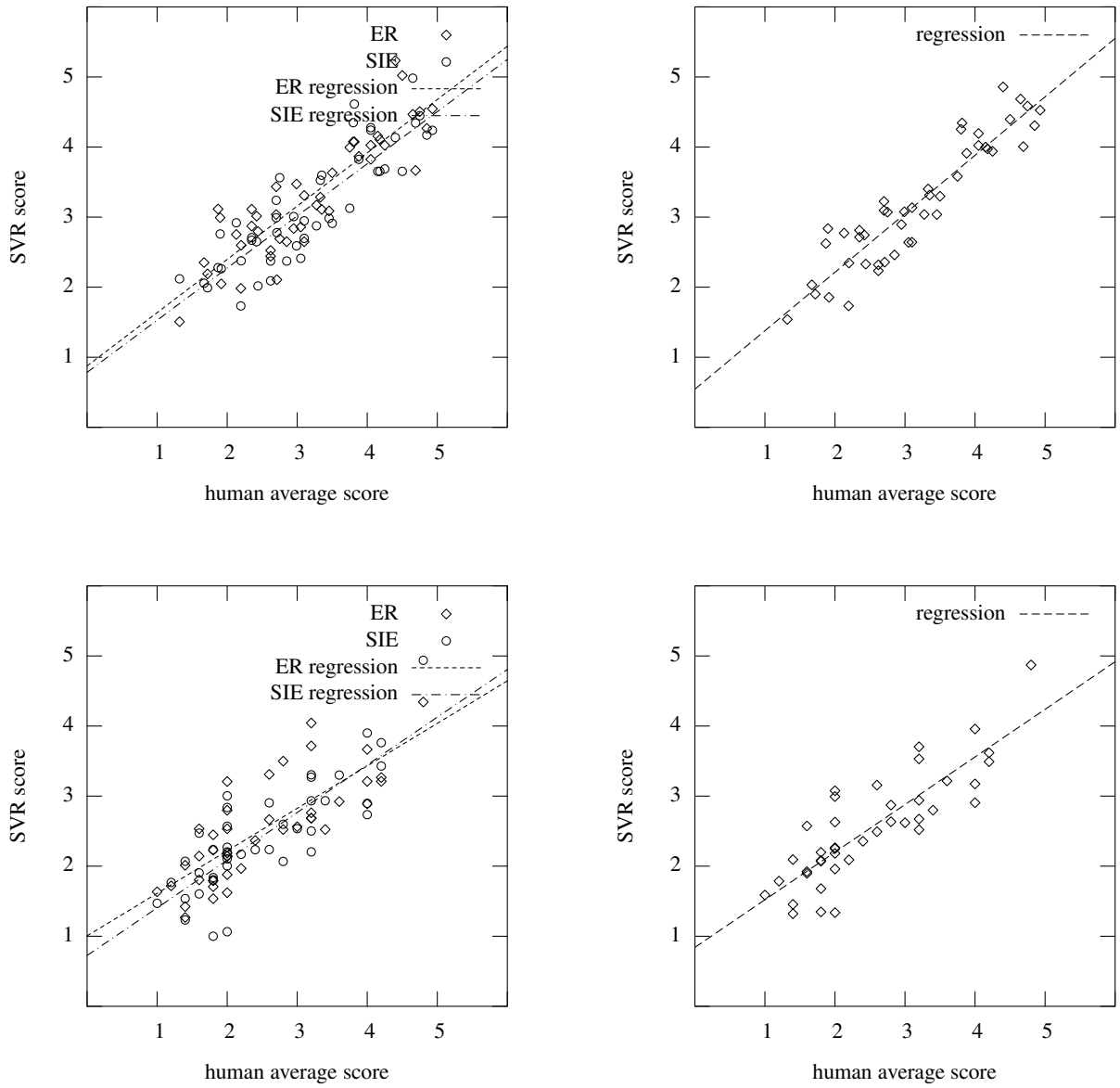

Fig. 1. The human average intelligibility score compared to the SVR-predicted score using the *ER* and *SIE* recognizer (left), and using the combination of both (right) for the *OC* (top) and *TE* data set (bottom).

For both data sets the ROVER technique did not show any valuable results in means of improving agreement between machine and human evaluation. However, experiments confirm that the word chain generated by ROVER using different recognizers can lead to very high WA and WR. This indicates that improving the recognition performance does not necessarily increase correlation if the spread of the WA and WR is reduced at the same time.

The first experiment was performed on the *OC* data set. Its signal quality is considered to be an upper border for real telephone signal quality. The WA of the *ER* and the WR of the *SIE* recognizer were combined using support vector regression to predict the human average intelligibility score. In a leave-one-out evaluation, the correlation could be increased to $r = 0.92$ ($\rho = 0.90$) which is the same as on the original data set *OC-orig* (see Table 2). Additionally, the average/maximum absolute errors could be reduced to 0.31/0.94 which is nearly the same as with the original data set (0.31/0.92, see Table 3).

The second experiment was performed on the *TE* data set. In contrast to the *OC* data set, the signal quality is considered to be a lower border for real telephone signal quality. The WA of the *ER* and *SIE* recognizer were combined in the same way as above. The correlation could be increased to $r = 0.84$ ($\rho = 0.84$) which comes

	Train = Test		LOO	
	$r(\mathrm{LR})$	$\rho(\mathrm{LR})$	$r(\mathrm{SVR})$	$\rho(\mathrm{SVR})$
SIE	0.82	0.82	0.79	0.81
ER	0.83	0.84	0.82	0.82
SIE+ER	**0.85**	**0.85**	**0.84**	**0.84**
ER/16 kHz	0.88	0.86	0.88	0.86

Table 5. Correlation of the WA of each single recognizer and the combination of the two values to the average human intelligibility rating using the *TE* data set. The ER/16 kHz results are based on the WA and the *TE-orig* data set. Note that the SVR results computed using LOO are nearly as good as the results using LR and a train = test scenario. (r – Pearson correlation coefficient, ρ – Spearman rank correlation coefficient)

close to the results of $r = 0.88$ ($\rho = 0.86$) using the *ER* recognizer on the original data set *TE-orig* (see Table 5). Additionally, the average/maximum absolute errors could be reduced to 0.42/1.09 which is a good step towards the results on the original data set (0.38/1.05, see Table 3).

The results presented in this work confirm our hypothesis that a combination of independent ASR systems can be used to recover the loss of the evaluation reliability due to the information lost in the speech signal during telephone transmission: experiments based on upper and lower border signal quality show very good results which are in the same range as those achieved with a single ASR system for 16 kHz close-talk microphone recordings. This is not only proved by means of agreement but also by the average and maximum absolute errors.

For further research, we suggest to explore the use of more than two different ASR systems for the evaluation of both telephone and close-talk speech as well as an evaluation of real telephone speech. It could also be interesting to evaluate other ways to combine two recognizers not yet considered in this paper, for instance using adaptation techniques described in [14]. The application to different test scenarios, i. e. scenarios depending on the number of words understood correctly instead of a Likert scale rating, needs to be investigated as well.

7. ACKNOWLEDGMENTS

This work was supported by *ELAN Fonds* of the *Universität Erlangen-Nürnberg* and the *Deutsche Krebshilfe* (German Cancer Aid) under grant 106266. The responsibility of the content of this paper lies with the authors.

8. REFERENCES

[1] R. Likert, "A Technique for the Measurement of Attitudes," *Archives of Psychology*, vol. 140, 1932.

[2] M. Schuster, E. Nöth, T. Haderlein, S. Steidl, A. Batliner, and F. Rosanowski, "Can you Understand him? Let's Look at his Word Accuracy — Automatic Evaluation of Tracheoesophageal Speech," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, USA, 2005, vol. 1, pp. 61–64, IEEE Computer Society Press.

[3] A. Maier, E. Nöth, E. Nkenke, and M. Schuster, "Automatic Assessment of Children's Speech with Cleft Lip and Palate," in *Proc. of the 5th Slovenian and 1st International Conference on Language Technologies (IS-LTC 2006)*, Ljubljana, Slovenia, 2006, pp. 31–35.

[4] A. Maier, M. Schuster, A. Batliner, E. Nöth, and E. Nkenke, "Automatic scoring of the intelligibility in patients with cancer of the oral cavity," in *Interspeech 2007 – Proc. Int. Conf. on Spoken Language Processing, 10th European Conference on Spoken Language Processing, 2007, Antwerp, Belgium*, 2007, to appear.

[5] K. Riedhammer, T. Haderlein, M. Schuster, F. Rosanowski, and E. Nöth, "Automatic Evaluation of Tracheoesophageal Telephone Speech," in *Proc. of the 5th Slovenian and 1st International Conference on Language Technologies (IS-LTC 2006)*, Ljubljana, Slovenia, 2006, pp. 17–22.

[6] G. Stemmer, *Modeling Variability in Speech Recognition*, vol. 19 of *Studien zur Mustererkennung*, Logos Verlag, Berlin, Germany, 2005.

[7] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, California, USA, 1997, pp. 347–352.

[8] I.H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques, 2nd Edition*, Morgan Kaufmann Pub, San Francisco, California, USA, 2005.

[9] Y. Wang and I.H. Witten, "Induction of model trees for predicting continuous classes," in *Proc. European Conference on Machine Learning*, Prague, Czech Republic, 1997.

[10] A.J. Smola and B. Schölkopf, "A tutorial on support vector regression," Royal Holloway College, University of London, UK, 1998.

[11] H. Liu and R. Setiono, "A probabilistic approach to feature selection - a filter solution," in *Proc. 13th International Conference on Machine Learning*, Bari, Italy, 1996, pp. 319–327.

[12] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, vol. 2, pp. 559–572, 1901.

[13] C. Spearman, "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, vol. 15, pp. 72–101, 1904.

[14] Diego Giuliani and Fabio Brugnara, "Acoustic model adaptation with multiple supervisions," in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006, pp. 151–154.