Influence of Reverberation on Automatic Evaluation of Intelligibility with Prosodic Features

Tino Haderlein[1], Michael Döllinger[2], Anne Schützenberger[2], and Elmar Nöth[1]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Lehrstuhl für Informatik 5 (Mustererkennung), Martensstraße 3, 91058 Erlangen, Germany
http://www5.cs.fau.de
Tino.Haderlein@fau.de

[2] Klinikum der Universität Erlangen-Nürnberg, Phoniatrische und pädaudiologische Abteilung in der HNO-Klinik, Bohlenplatz 21, 91054 Erlangen, Germany

Abstract. Objective analysis of intelligibility by a speech recognizer and prosodic features was performed for close-talking recordings before. This study examined whether this is also possible for reverberated speech. In order to ensure that only the room acoustics are different, artificial reverberation was used. 82 patients after partial laryngectomy read a standardized text, 5 experienced raters assessed intelligibility perceptually on a 5-point scale. The best feature subset, determined by Support Vector Regression, consists of the word correctness of a speech recognizer, the average duration of silent pauses, the standard deviation of the F_0 on the entire sample, the standard deviation of jitter, and the ratio of the durations of the voiced sections and the entire recording. A human-machine correlation of $r = 0.80$ was achieved for the close-talking recordings and $r = 0.72$ for the worst case of the examined signal qualities. By adding three more features, also $r = 0.80$ was reached for the reverberated scenario.

Keywords: intelligibility, automatic assessment, prosody, SVR, reverberation

1 Introduction

Automatic evaluation of voice and speech impairment allows to obtain objective assessment of the current state and temporal changes of distorted communication by voice [1]. In this study, the topic is the evaluation of intelligibility after partial removal of the larynx. Earlier work has shown that this can be performed using an automatic speech recognition (ASR) system, supported by a prosodic analysis module [2]. As a reference evaluation, the perceptual assessment by a speech therapist is the standard.

In order to get the best possible acoustic quality of the recordings to be analyzed, usually a headset or another close-talking microphone is used. However, this recording situation might have a negative influence on the patient. The patient might feel watched or controlled when he or she is aware that other people could get access to the recording. For patients after head or neck surgery, wearing a headset can also be painful. If the microphone is somewhere else in the room, both effects are attenuated, but the samples will be affected by reverberation then. It has been shown that, with according training data for the ASR system, speech recognition does also work in reverberated environment, even if the properties in the recording environment, such as the room impulse

response, are not exactly known [3]. It has, however, not been examined whether a close-talking ASR system, supported by prosodic analysis, can be also used for intelligibility assessment in reverberated environment, i.e. for larger distances between mouth and microphone, and whether the same prosodic features are optimal for different distances. This will be addressed in this paper. The reference evaluation for all scenarios, however, is the human evaluation of a close-talking speech sample, because in a therapy session the patient and the therapist sit close to each other.

This paper is organized as follows: Section 2 introduces the speech data used for the experiments, Sect. 3 describes the artificial reverberation of these data. The processing of the features from the speech recognizer (Sect. 4) and the prosody module (Sect. 5) by Support Vector Regression follows in Sect. 6. The results will be discussed in Sect. 7.

2 Test Data and Subjective Evaluation

Two sets of speech data have been used for this study. The samples of the first set, further denoted as set *t82*, were used to determine prosodic measures of reverberated recordings that can describe perceptual intelligibility results. The second set – here called set *t85* – was used to test whether the findings on set *t82* also hold for other data with the same type of voice disorder. Set *t82* contains recordings of 82 persons (68 men and 14 women) after partial laryngectomy due to laryngeal cancer. Their average age was 62.3 with a standard deviation of 8.8 years; the youngest speaker was 41, the oldest one was 86 years old. Set *t85* was recorded from 85 patients (75 men, 10 women) suffering from cancer in different regions of the larynx. 65 of them had already undergone partial laryngectomy, 20 speakers were still awaiting surgery. The average age of the speakers was 60.7 years with a standard deviation of 9.7 years. The youngest and the oldest person were 34 and 83 years old, respectively.

Informed consent had been obtained by all participants prior to the examination. The study respected the principles of the World Medical Association (WMA) Declaration of Helsinki on ethical principles for medical research involving human subjects and has been approved by the ethics committee of our university. All persons read the German version of the tale “The North Wind and the Sun” [4], which is widely used in medical speech evaluation in German-speaking and other countries. It consists of 71 disjoint words and 108 words in total (172 syllables). The patients were recorded by a close-talking microphone (Logitech Premium Stereo Headset 980369-0914) with 16 kHz sampling frequency and 16 bit linear amplitude resolution.

Five experienced voice professionals (ear-nose-throat doctors, speech therapists) evaluated the intelligibility of each recording. The samples were played to the experts once via loudspeakers in a quiet seminar room without disturbing noise or echoes. Rating was performed on a five-point Likert scale. For computation of average scores for each patient, the grades were converted to integer values (1 = ‘very high’, 2 = ‘rather high’, 3 = ‘medium’, 4 = ‘rather low’, 5 = ‘very low’). The sets *t82* and *t85* were evaluated by two different rater groups. Three of the therapists were part of both groups; however, set *t85* had already been evaluated about one year before set *t82*. The human raters evaluated only the original close-talking recordings and not the artificially reverberated samples that will be introduced in Sect. 3.

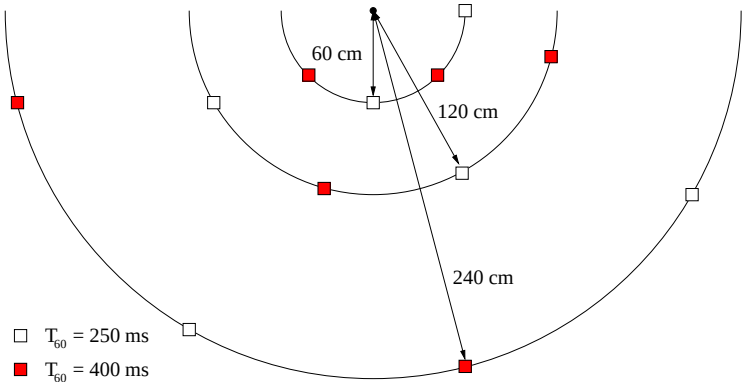

Fig. 1. Measuring room impulse responses for artificial reverberation (microphone position: •) at 12 speaker positions in a room with variable reverberation times T_{60}

3 Artificial Reverberation of Speech Samples

In order to obtain reverberated speech with a large variety of acoustic quality, speech data must be collected in many rooms with different impulse responses. This leads to the problem that not always the same speakers are available, and if so, they will not be able to reproduce a given text exactly as in the other recording sessions. Hence, the recordings will not only be different with respect to the room impulse response, but also with respect to speaking style and maybe vocabulary. Reverberating close-talking speech artificially with previously measured room impulse responses can avoid this problem and also drastically reduce the effort of data acquisition.

For this study, the required impulse responses were obtained in a room where the reverberation time could be changed from $T_{60} = 250$ ms to $T_{60} = 400$ ms by removing sound absorbing carpets and curtains. 12 impulse responses were measured for loudspeaker positions on three semi-circles in front of the microphone at distances 60 cm, 120 cm, and 240 cm (see Fig. 1 and Table 2). The recording angle was counted clockwise from 0° to 165°. Six impulse responses each were measured with $T_{60} = 250$ ms and $T_{60} = 400$ ms. The available close-talking speech data (Sect. 2) were reverberated with each of them so that 12 reverberated versions of the original samples were available.

4 The Speech Recognition System

The speech recognition system used for the experiments has been described in [5]. It is based on semi-continuous Hidden Markov Models (HMM) and was trained with close-talking speech only. For each 16 ms frame, a 24-dimensional feature vector is computed. It contains short-time energy, 11 Mel-frequency cepstral coefficients, and the first-order derivatives of these 12 static features. The recognition vocabulary of the recognizer was changed to the 71 words of the standard text. The word accuracy and the word correctness were used as basic automatic measures for intelligibility since they

had been successful for other voice and speech pathologies [6, 7]. They are computed from the comparison between the recognized word sequence and the reference text consisting of the $n_{\mathrm{all}} = 108$ words of the read text. With the number of words that were wrongly substituted (n_{sub}), deleted (n_{del}) and inserted (n_{ins}) by the recognizer, the word accuracy in percent is given as

$$\mathrm{WA} = [1 - (n_{\mathrm{sub}} + n_{\mathrm{del}} + n_{\mathrm{ins}})/n_{\mathrm{all}}] \cdot 100$$

while the word correctness omits the wrongly inserted words:

$$\mathrm{WR} = [1 - (n_{\mathrm{sub}} + n_{\mathrm{del}})/n_{\mathrm{all}}] \cdot 100$$

Only a unigram language model was used so that the results mainly depend on the acoustic models. A higher-order model would correct too many recognition errors and thus make WA and WR useless as measures for intelligibility.

5 Prosodic Features

In order to find automatically computable counterparts for intelligibility, also a 'prosody module' was used to compute features based upon frequency, duration, and speech energy (intensity) measures. This is common in automatic speech analysis on normal voices [8–10]. The prosody module processes the output of the word recognition module and the speech signal itself. 'Local' prosodic features are computed for each word position. Originally, there were 95 of them. After several studies on voice and speech assessment, however, a relevant core set of 33 features has been defined for further processing [11]. The components of their abbreviated names are given in parentheses:

- Length of pauses (Pause): length of silent pause before (–before) and after (–after), and filled pause before (Fill-before) and after (Fill-after) the respective word
- Energy features (En): regression coefficient (RegCoeff) and the mean square error (MseReg) of the energy curve with respect to the regression curve; mean (Mean) and maximum energy (Max) with its position on the time axis (MaxPos); absolute (Abs) and normalized (Norm) energy values
- Duration features (Dur): absolute (Abs) and normalized (Norm) duration
- F_0 features (F0): regression coefficient (RegCoeff) and mean square error (MseReg) of the F_0 curve with respect to its regression curve; mean (Mean), maximum (Max), minimum (Min), voice onset (On), and offset (Off) values as well as the position of Max (MaxPos), Min (MinPos), On (OnPos), and Off (OffPos) on the time axis; all F_0 values are normalized.

The last part of the feature name denotes the context size, i.e. the interval of words on which the features are computed (see Table 1). They can be computed on the current word (W) or in the interval that contains the second and first word before the current word and the pause between them (WPW). A full description of the features used is beyond the scope of this paper; details and further references are given in [5, 12].

Besides the 33 local features per word, 16 'global' features were computed for intervals of 15 words length each. They were derived from jitter, shimmer, and the number

of detected voiced and unvoiced sections in the speech signal [12]. They covered the means and standard deviations of jitter and shimmer, the number, length, and maximum length of voiced and unvoiced sections, the ratio of the numbers of voiced and unvoiced sections, the ratio of the length of the voiced sections to the length of the signal, and the same for unvoiced sections. The standard deviation of the F_0 was measured in two ways: it was computed for all voiced sections only and also over all sections of the speech recordings. In the latter case, each unvoiced frame contributed a value of 0. Hence, it incorporated also information about the percentage of frames where no regular voice signal was detected. Since all patients read the same text, this was supposed to indicate the degree of pathology.

The human listeners gave ratings for the entire text. In order to receive also one single value for each feature that could be compared to the human ratings, the average of each prosodic feature over the entire recording served as final feature value.

Table 1. Local prosodic features; the context size denotes the interval of words on which the features are computed (W: one word, WPW: word-pause-word interval)

features	context size	
	WPW	W
Pause: before, Fill-before, after, Fill-after		•
En: RegCoeff, MseReg, Abs, Norm, Mean	•	•
En: Max, MaxPos		•
Dur: Abs, Norm	•	•
F0: RegCoeff, MseReg	•	•
F0: Mean, Max, MaxPos, Min, MinPos, Off, OffPos, On, OnPos		•

6 Support Vector Regression (SVR)

In order to determine the best subset of WA, WR, and prosodic features to model the human intelligibility rating, Support Vector Regression (SVR, [13]) was used. The underlying SVM used a linear kernel. The complexity constant C for the SVR was set to 1. Each training example for the regression consisted of a set of features from the close-talking *t82* data set (the inputs) and a human intelligibility score (the target output). The sequential minimal optimization algorithm (SMO, [13]) of the Weka toolbox [14] was applied in a 10-fold cross-validation manner. For the attribute selection, the Greedy Stepwise algorithm was applied. The standard settings were not changed. In contrast to [2], all input features were not normalized but standardized (mean value: $\mu = 0$, standard deviation: $\sigma = 1$) for the analysis.

7 Results and Discussion

Former experiments on data set *t82* had revealed optimal features for close-talking analysis [2]. The human-machine correlation was $r = 0.79$ when prosodic features and the

word correctness (WR) were combined. Using the WA instead gave worse results. The best feature subset consisted of WR, the average duration of the silent pauses before a word (Pause–before), the standard deviation of the fundamental frequency on the entire sample (StandDevF0/Sig), the standard deviation of jitter (StandDevJitter), and the ratio of the durations of the voiced sections and the entire recording (RelDur+Voiced/Sig). This set revealed also the best results on the same data in this study ($r = 0.80$; see Table 2), although the setup of the regression has been slightly changed.

In earlier experiments on the analysis of chronic hoarseness, also the absolute energy measured in a word-pause-word interval (EnAbsWPW) was among the candidates for the best feature set [15]. When this feature was added in this study, however, no significant rise of human-machine correlations could be obtained, except for impulse response `h421045` ($r = 0.78$ instead of 0.77) and `h423075` ($r = 0.74$ instead of 0.72). All other correlations were equal to those of the smaller set or even lower. For this reason, EnAbsWPW has been removed from the feature set again.

As expected, the human-machine correlation got lower in a higher reverberation time and with rising microphone distance, but still the lowest measured correlation was as high as 0.72 for 240 cm microphone distance and $T_{60} = 400$ ms where it had been 0.80 for the close-talking case. The angle at which the speaker spoke towards the microphone did not show consistent influence. The standard situation would be 90°, i.e. right in front of the microphone. All available pairs of the same T_{60} and microphone distance showed a relative difference in the angle of 90°, but very often not the one, which was closer to the absolute 90° position, was better with respect to human-machine correlation. Additionally, the influence of the angle was usually only $\Delta r = 0.02$.

The weights of the single features in the regression formulae for the *t82* data (see Table 2) do not show a unique behavior that would make it easy to relate them to T_{60} or microphone distance. The weight for the pauses between words (Pause–before) is relatively stable at about 0.2 to 0.4 among the simulated recording situations. The standard deviation of jitter (StandDevJitter) tends to get higher absolute values for more reverberant environments. The ratio of the durations of the voiced segments and the whole recording (RelDur+Voiced/Sig) loses influence with rising reverberation. So does also the word correctness (WR). The reason is obviously the mismatch between the acoustic properties of training and test environment of the recognizer. For the F_0 and jitter features, also the worse acoustic quality and hence the unreliable detection of F_0 is the most probable reason. However, the weight for StandDevF0/Sig is at about the same level in all experiments, and the StandDevF0, which is not related to the overall duration, but only to the voiced segments, does not occur in the best feature sets at all.

In general, the feature set, which was best for the close-talking case, can also basically be considered suitable for reverberated environment. In another experiment, a feature set has been determined, which is best for the acoustic scenario most deviant from the close-talking case (impulse response `h423075`, $T_{60} = 400$ ms, microphone distance 240 cm, angle 165°; see Table 3). Here, for the *t82* data the best feature set is a superset of the one for the close-talking case. It contains additionally the mean of jitter, the number of voiced sections in the recording (#+Voiced), and the ratio of the numbers of voiced and unvoiced sections (RelNum+/–Voiced). This lifts the human-machine correlation from 0.74 to 0.80. For the close-talking recordings, however, the correlation drops from 0.80 to 0.73 with this set, and the additional features show very low regres-

sion weights. The results on the *t85* data set (Table 2 and 3) confirm the suitability of the selected features for intelligibility assessment both in good and bad acoustic conditions.

Table 2. Feature weights (columns 5 to 9) and human machine correlation r for artificially reverberated and close-talking *t82* data (last line: *t85*); reverberation time T_{60}, microphone distance ('dist.') and recording angle for the impulse responses are given on the left side

impulse response	T_{60} (ms)	dist. (cm)	angle (°)	Pause–before	StandDev Jitter	RelDur+ Voiced/Sig	WR	StandDev F0/Sig	r
h411000	250	60	0	0.207	0.074	–0.542	–0.419	0.524	0.78
h411090	250	60	90	0.253	0.151	–0.676	–0.356	0.521	0.79
h412060	250	120	60	0.313	0.136	–0.699	–0.377	0.468	0.78
h412150	250	120	150	0.256	–0.067	–0.593	–0.323	0.560	0.80
h413030	250	240	30	0.260	–0.234	–0.436	–0.376	0.573	0.77
h413120	250	240	120	0.238	–0.194	–0.451	–0.304	0.511	0.75
h421045	400	60	45	0.286	0.179	–0.700	–0.382	0.471	0.77
h421135	400	60	135	0.245	0.225	–0.782	–0.362	0.544	0.80
h422015	400	120	15	0.232	–0.265	–0.414	–0.154	0.524	0.76
h422105	400	120	105	0.322	–0.174	–0.471	–0.267	0.549	0.74
h423075	400	240	75	0.387	–0.222	–0.328	–0.116	0.472	0.72
h423165	400	240	165	0.292	–0.325	–0.364	–0.237	0.570	0.74
– (close-talk)	—	3–5	90	0.191	0.223	–0.881	–0.412	0.511	0.80
– (close-t., t85)	—	*3–5*	*90*	*0.485*	*–0.013*	*–0.551*	*–0.313*	*0.554*	*0.73*

Table 3. Feature weights and human-machine correlation r for the set optimal for reverberated (impulse response h423165) *t82* data, tested on reverberated and close-talking *t82* and *t85* data

impulse response	Pause–before	Mean Jitter	StandDev Jitter	#+Voiced	RelNum +/–Voiced	RelDur+ Voiced/Sig	StandDev F0/Sig	r
h423165 *(t82)*	0.290	0.895	–0.970	–0.422	–0.206	–0.200	0.608	0.80
h423165 (t85)	*0.504*	*0.640*	*–0.553*	*–0.121*	*–0.349*	*–0.292*	*0.452*	*0.75*
– (close-t., *t82*)	0.432	0.027	0.147	–0.072	0.056	–0.847	0.508	0.73
– (close-t., t85)	*0.377*	*0.658*	*–0.659*	*–0.215*	*–0.288*	*–0.403*	*0.493*	*0.74*

The inter-rater correlation between a single rater's intelligibility scores and the average of the 4 other raters was $r = 0.84$ for the *t82* recordings (for details, see [2]) and $r = 0.81$ for the *t85* data [16]. These are the reference values that an automatic system should reach to be regarded as reliable as an average human rater. The current results are almost at this level. Due to slight differences in correlations and regression weights for different features that occurred in this study, the experiments have to be continued with larger data sets in order to reassure the relevance of the selected features and to

add other features where applicable. Nevertheless, the conclusion of this study is that automatic evaluation of intelligibility can be done on reverberated speech samples as reliable as for close-talking samples.

Acknowledgments: We would like to thank Dr. Wolfgang Herbordt for his kind support with the software and data for artificial reverberation. Dr. Döllinger's contribution was supported by Deutsche Krebshilfe grant no. 111332.

References

1. Baghai-Ravary, L., Beet, S.: Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders. Springer, New York (2013)
2. Haderlein, T., Nöth, E., Batliner, A., Eysholdt, U., Rosanowski, F.: Automatic Intelligibility Assessment of Pathologic Speech over the Telephone. Logoped. Phoniatr. Vocol. **36** (2011) 175–181
3. Couvreur, L., Couvreur, C., Ris, C.: A Corpus-Based Approach for Robust ASR in Reverberant Environments. In: Proc. ICSLP. Volume 1, Beijing (2000) 397–400
4. International Phonetic Association (IPA): Handbook of the International Phonetic Association. Cambridge University Press, Cambridge (1999)
5. Haderlein, T., Moers, C., Möbius, B., Rosanowski, F., Nöth, E.: Intelligibility Rating with Automatic Speech Recognition, Prosodic, and Cepstral Evaluation. In Habernal, I., Matoušek, V., eds.: Proc. TSD 2011. Volume 6836 of LNAI., Berlin, Heidelberg, Springer (2011) 195–202
6. Haderlein, T.: Automatic Evaluation of Tracheoesophageal Substitute Voices. Volume 25 of Studien zur Mustererkennung. Logos Verlag, Berlin (2007)
7. Maier, A.: Speech of Children with Cleft Lip and Palate: Automatic Assessment. Volume 29 of Studien zur Mustererkennung. Logos Verlag, Berlin (2009)
8. Nöth, E., Batliner, A., Kießling, A., Kompe, R., Niemann, H.: VERBMOBIL: The Use of Prosody in the Linguistic Components of a Speech Understanding System. IEEE Trans. on Speech and Audio Processing **8** (2000) 519–532
9. Rosenberg, A.: Automatic Detection and Classification of Prosodic Events. PhD thesis, Columbia University, New York (2009)
10. Origlia, A., Alfano, I.: Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabification. In Calzolari, N., et al., eds.: Proc. 8th Int. Conf. on Language Resources and Evaluation (LREC'12). (2012) 997–1002
11. Haderlein, T., Schwemmle, C., Döllinger, M., Matoušek, V., Ptok, M., Nöth, E.: Automatic Evaluation of Voice Quality Using Text-based Laryngograph Measurements and Prosodic Analysis. Comput. Math. Methods. Med. **2015** (2015) 11 pages. Published June 2, 2015.
12. Batliner, A., Buckow, J., Niemann, H., Nöth, E., Warnke, V.: The Prosody Module. In Wahlster, W., ed.: Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin (2000) 106–121
13. Smola, A.J., Schölkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing **14** (2004) 199–222
14. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd edn. Morgan Kaufmann, San Francisco (2005)
15. Haderlein, T., Döllinger, M., Matoušek, V., Nöth, E.: Objective Voice and Speech Analysis of Persons with Chronic Hoarseness by Prosodic Analysis of Speech Samples. Logoped. Phoniatr. Vocol. (2015) Epub 2015 May 27.
16. Bocklet, T., Haderlein, T., Hönig, F., Rosanowski, F., Nöth, E.: Evaluation and Assessment of Speech Intelligibility on Pathologic Voices Based upon Acoustic Speaker Models. In Godino-Llorente, J., Gómez-Vilda, P., eds.: 3rd Advanced Voice Function Assessment International Workshop (AVFA2009), Madrid, Universidad Politécnica de Madrid (2009) 89–92