published in: Text, Speech and Dialogue (TSD 2018), Proc.; LNAI 11107, ©Springer Nature Switzerland, 2018, pp. 473–481; www.springer.com/series/1244

Subtext Word Accuracy and Prosodic Features for Automatic Intelligibility Assessment

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

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

[2] Universitätsklinikum Erlangen, Phoniatrische und Pädaudiologische Abteilung in der
HNO-Klinik, Waldstraße 1, 91054 Erlangen, Germany

Abstract. Speech intelligibility for voice rehabilitation can successfully be evaluated by automatic prosodic analysis. In this paper, the influence of reading errors and the selection of certain words (nouns only, nouns and verbs, beginning of each sentence, beginnings of sentences and subclauses) for the computation of the word accuracy (WA) and prosodic features are examined. 73 hoarse patients read the German version of the text "The North Wind and the Sun". Their intelligibility was evaluated perceptually by 5 trained experts according to a 5-point scale. Combining prosodic features and WA by Support Vector Regression showed human-machine correlations of up to $r = 0.86$. They drop for files with few reading errors, however, but this can largely be evened out by feature set adjustment. WA should be computed on the whole text, but for some prosodic features, a subset of words may be sufficient.

Keywords: intelligibility, automatic assessment, prosody, reading errors

1 Introduction

Established methods for objective voice and speech evaluation in therapy analyze only sustained vowels. In our approach for the assessment of speech intelligibility, the test persons read a given standard text that undergoes prosodic analysis. Usually, each prosodic feature has been averaged over all words in the text. However, it is widely known that intelligibility varies among different word classes which is mostly caused by prosodic properties [1–5]. Hence, putting all content and function words, long and short words, and words at different positions in sentences, together bears the risk of losing information. In previous work [6], the influence of the position and type of words, which are selected from a read-out text, on the reliability of the automatic analysis has already been addressed. However, this was restricted to single prosodic features. The word accuracy (WA) of a speech recognizer has also been used as basic measure for intelligibility [7, 8]. It has also always been computed for an entire text sample. In this follow-up study, the suitability of the WA computed on subunits of a text will be examined, and the combination of these features and prosodic features by Support Vector Regression will be presented for the first time. It has further been shown that the automatic analysis is influenced by the number of reading errors in the sample [6]. This will be tested for the new feature sets. These main questions are addressed in this paper:

- How can prosodic features and the word accuracy together model the human intelligibility rating when they are computed on different subparts of a standard text?
- Does the number of reading errors in a speech sample have an influence on the human and automatic intelligibility rating?

This work is organized as follows: Section 2 introduces the test data and the perceptual evaluation reference. The computation of the features is described in Sect. 3. The results of the experiments (Sect. 4) will be discussed in Sect. 5.

2 Test Data and Subjective Evaluation

73 German subjects with chronic hoarseness participated in this study (Table 1). Patients suffering from cancer were excluded. Each person read the text "Der Nordwind und die Sonne" ("The North Wind and the Sun", [9]), a phonetically rich standard text which is frequently used in clinical speech evaluation in German-speaking countries. It contains 108 words (71 distinct) with 172 syllables. The data were recorded with a sampling frequency of 16 kHz and 16 bit amplitude resolution using an AKG C 420 microphone (AKG Acoustics, Vienna, Austria). They were recorded in a quiet room in our university and digitally stored on a server by a client/server-based system [10, Chap. 4]. 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 clinics.

Five voice professionals (one ear-nose-throat doctor, four speech therapists) evaluated the intelligibility of each original recording perceptually. 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'). For each patient, an intelligibility mark, expressed as a floating point value, was calculated as the arithmetic mean of the single scores. These marks served as ground truth in our experiments.

Table 1. The test speakers (entire set, group with few and group with many reading errors)

group	persons			age				reading errors			
	all	men	women	μ	σ	min	max	μ	σ	min	max
overall	73	24	49	48.3	16.8	19	85	3.10	3.50	0	17
low-error	32	9	23	48.5	13.7	26	76	0.34	0.47	0	1
high-error	41	15	26	48.1	18.9	19	85	5.24	3.34	2	17

Due to reading errors, repetitions, and additional remarks, such as "read now?", the recordings did not only contain words appearing in the text reference but also additional words and word fragments. In order to describe the errors, a manual word-based counting of errors was adopted (see details in [6, 7]). In order to study the effect of errors on the evaluation on subsets of reasonable size, the overall data set was divided into a 'low-error' group with at most one reading error per speaker and a 'high-error' group with 2 to 17 errors per speaker (Table 1). The left side of Fig. 1 shows that there are also low-error readers with a bad perceptual ranking. The human intelligibility rating and the number of reading errors are just weakly correlated ($r = 0.35$).

3 Prosodic Features

The speech recognizer used for the experiments [11] is based on semi-continuous Hidden Markov Models (HMM). 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. Only a unigram language model was used so that the results mainly depend on the acoustic models.

In order to find counterparts for intelligibility, a 'prosody module' was used to compute features based upon frequency, duration, and speech energy (intensity) measures. 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 [12]. 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 2). 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 [11, 13].

Besides the 33 local features per word, 15 '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 [13]. 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 last feature was the standard deviation of the F_0.

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 all selected words served as final feature value.

Table 2. 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		•

4 Experiments

Earlier experiments averaged each prosodic feature over the entire read-out text. For this study, we examined whether the restriction to certain subsets might be beneficial:

- averaging over *all words* (the baseline; 108 words, denoted by the suffix '_all')
- *nouns only* (24 words, '_nouns')
- *nouns and verbs* (44 words, '_n+v')
- *beginnings of sentences* (first 3 words of each of the 6 sentences; 18 words, '_sent_i')
- *beginnings of sentences and subclauses* (first 3 words of each of the 6 sentences and 10 subclauses; 48 words, '_s+s_i')

Nouns and verbs were chosen because content words generally show less predictability and hence intelligibility than function words, such as articles, prepositions, and conjunctions [14]. Adjectives were not taken into account because there are very few in the text. These words contribute to intelligibility mainly because of their stress patterns, one of the main aspects the prosodic features were designed for. The beginnings of sentences and subclauses, without the regard of the word classes, were chosen with respect to the medical application. Many voice and speech patients show higher speaking effort and shorter phonation time, so they will have to pause more often and fragment the paragraph to be read. Breaks usually occur at syntactic boundaries.

In former studies [7, 8], the word accuracy (WA) of a speech recognizer was an important feature to model intelligibility. It is computed from the comparison of the recognized word sequence and the reference text consisting of the $n_{\rm all} = 108$ words of the read text. With the number of words that were wrongly substituted ($n_{\rm sub}$), deleted ($n_{\rm del}$) and inserted ($n_{\rm ins}$) by the recognizer, the word accuracy in percent is given as

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

In this study, it is also added to the feature set, but just like the prosodic features, several versions of it will be used for the first time: besides WA_all, computed on all words of the text, there is also WA_nouns, WA_n+v, WA_sent_i, and WA_s+s_i.

In order to find the best subset of WA and the prosodic features to model the subjective intelligibility ratings, Support Vector Regression (SVR) was applied. For this study, the sequential minimal optimization algorithm [15] of Weka [16] was applied. Due to the small amount of available data, a 10-fold cross-validation was used. For the regression, the automatically computed measures (WA and all prosodic features) served as the training set. The test set consisted of the subjective, perceptual intelligibility scores. Due to the small amount of data, we also refrained from using deep learning technology.

5 Results and Discussion

The first experiment was performed using the different word accuracy types as a single feature. The human-machine correlation showed the best results for WA_all ($r = –0.74$, first part of Table 3). The other values are – in some cases substantially – lower.

The next part of the table shows a comparison of the correlations for the best single prosodic features, as determined in [6], and the predicted intelligibility of the SVR when all prosodic features were put together. The single features could compete with the WA results, the combination of all prosodic features achieves substantially better values. The best result was $r = 0.84$ for all features being averaged over all words.

The next lines of Table 3 show the use of prosodic features from a subtext scenario x supplemented either by WA_all or WA_x. These results are even better than for the prosodic features alone. In general, WA_all contributes better to the human-machine agreement than the WA from the other scenarios (with two non-significant exceptions). WA_all and prosodic features obtained on the whole text achieved $r = 0.86$. The right side of Fig. 1 shows this case. The ‘real’ ratings can only be between 1 and 5. If the predicted values outside that range are replaced by the possible minimum and maximum values (i. e. in our data 0.909 by 1, 5.714 and 6.023 by 5), the correlation does not improve, however. Without the outlier at position (0.909, 2.400), the correlation for the low-error files would rise from 0.74 to 0.80.

Computing the prosodic features on a lower number of words keeps the human-machine agreement at a high level in general, with the best result for sentence and subunit beginnings (_s+s_i) where only a drop in correlation of $\Delta r = 0.01$ is measured.

Next, the influence of the reading errors on the combined feature set was examined. The low-error files show significantly lower correlations, just like for single features [6], while for the high-error files, the results stay at about $r = 0.80$ if WA_all is used.

Table 4 contains the regression weights for the case that all 73 speech samples are used in the SVR. The baseline is given in the first data column, i. e. using WA_all and the prosodic features computed on all words. WA_all shows a consistently high weight in all setups. The other WA_x for the respective scenarios x show lower weights on the average. WA_nouns is not even part of the best set for the whole database. The same holds for WA_sent_i and WA_s+s_i on the low-error files. Like in a similar study with a slightly different SVR setup on the _all scenario [8], also the duration feature DurNormWPW, representing the speaking rate, becomes important in some cases. Further, the F_0 value at voice offset (F0OffW) appears in some sets. It likely resembles voice quality and stability. Similar information may be inherent in #+Voiced, i. e. the number of voiced segments, that appears in the sets for the _sent_i and _s+s_i cases (when using WA_all).

The new results mostly confirm the earlier study [6] that identified single prosodic features with a high human-machine correlation. Pause–before and the regression coefficient of the energy in a word-pause-word interval (EnRegCoeffWPW) were in the new tests not among the best sets, however. The normalized duration of a word-pause-word interval (DurNormWPW) has also been a good indicator for intelligibility in earlier studies and could mostly replace the energy EnNormWPW [7]. Here, it plays only a minor role. Both DurNormWPW and Pause–before reveal the overall speaking rate.

EnNormWPW is a good indicator for intelligibility [6, 7, 11]. Especially for low-error reading, a selection of words from the text lowers the correlation to the perceptual scores, however. MeanJitter showed the highest correlation of all in [6], namely $r = 0.73$ for low-error reading and the _n+v case. In this new study, it is present in all scenarios.

In a final experiment, all features from all scenarios were combined. The best subset comprised EnNormWPW_all, F0OffW_all, MeanJitter_n+v, and WA_all, achieving $r = 0.86$ for the whole data set. In this case only the MeanJitter_n+v had to be computed for only 44 words and the other features for all 108 words. So the benefit of restricting the word set for computation seems too small to be useful. However, with these features, the correlation was $r = 0.73$ for the low-error and $r = 0.87$ for the high-error files. With MeanJitter_all, only $r = 0.59$ had been measured on the low-error files, so the use of mixed scenarios (_all, _n+v) shows on the average the most stable results for all tested cases.

We are aware of the problem arising when standard texts are used for measuring intelligibility. However, it was shown that text-based evaluation performed by trained listeners is as reliable as an inverse intelligibility test, where naïve raters write down a previously unknown word sequence read by the test person. For more details, see [6].

As a conclusion, it can be stated that the word accuracy (WA) should always be used in a feature set for intelligibility assessment, and it should be computed on the full text while it is sufficient to compute some of the prosodic features only on the first three words of a sentence and subclause. Only WA_n+v and the _n+v prosodic features give a slightly better result than WA_all and _n+v features for low-error files. The _n+v case already showed the best results for single prosodic features [6]. The influence of many reading errors is a positive one at first sight since the human-machine correlation is better for samples with many errors. However, Fig. 1 (left) shows that the low-error files are concentrated on a much smaller perceptual range than the high-error files. Hence, it is more difficult to find a feature set mapping these small differences. More effort has to be put on this in the next experiments. Future work also includes tuning of the SVR parameters. Preliminary tests changing the kernel parameter C in a range of 0.01 to 1000 have shown no better results, however.

Acknowledgments: Dr. Döllinger's contribution was supported by the German Research Foundation (DFG), grant no. DO1247/8-1 (no. 323308998).

References

1. Hustad, K., Dardis, C., McCourt, K.: Effects of visual information on intelligibility of open and closed class words in predictable sentences produced by speakers with dysarthria. Clin

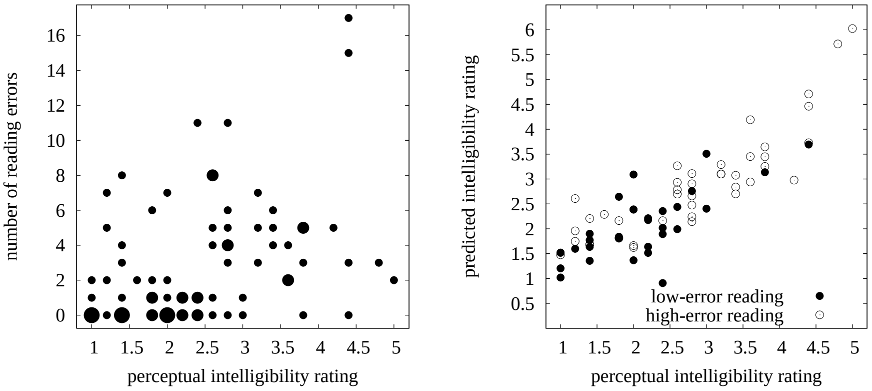

Fig. 1. Left side: average human intelligibility rating vs. number of reading errors in the speech files; the point size indicates single, double, or triple occurrence. Right side: average human vs. predicted intelligibility rating for WA_all and the _all prosodic features

Linguist Phon **21** (2007) 353–367

2. Cutler, A.: Phonological cues to open- and closed-class words in the processing of spoken sentences. J Psycholinguist Res **22** (1993) 109–131
3. Grosjean, F., Gee, J.: Prosodic structure and spoken word recognition. Cognition **25** (1987) 135–155
4. Pichney, M., Durlach, N., Braida, L.: Speaking clearly for the hard of hearing. II: Acoustic characteristics of clear and conversational speech. J Speech Hear Res **29** (1986) 434–446
5. Turner, G., Tjaden, K.: Acoustic differences between content and function words in amyotrophic lateral sclerosis. J Speech Lang Hear Res **43** (2000) 769–781
6. Haderlein, T., Schützenberger, A., Döllinger, M., Nöth, E.: Robust Automatic Evaluation of Intelligibility in Voice Rehabilitation Using Prosodic Analysis. In Ekštein, K., Matoušek, V., eds.: Proc. TSD 2017. Volume 10415 of LNAI. Cham, Switzerland, Springer International Publishing Switzerland (2017) 11–19
7. Haderlein, T., Nöth, E., Maier, A., Schuster, M., Rosanowski, F.: Influence of Reading Errors on the Text-Based Automatic Evaluation of Pathologic Voices. In Sojka, P., Horák, A., Kopeček, I., Pala, K., eds.: Proc. TSD 2008. Volume 5246 of LNAI. Berlin, Heidelberg, Springer (2008) 325–332
8. 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 **41** (2016) 106–116
9. International Phonetic Association (IPA): Handbook of the International Phonetic Association. Cambridge University Press, Cambridge (1999)
10. Maier, A.: Speech of Children with Cleft Lip and Palate: Automatic Assessment. Volume 29 of Studien zur Mustererkennung. Logos Verlag, Berlin (2009)
11. 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

Table 3. Human-machine correlation r for SVR values obtained from prosodic features and WA on all 73 speech samples, depending on the scenario x used for computation; bold-face: best results of each line; italics: WA_*x* not part of best feature set

	scenario *x*	all	nouns	n+v	sent_i	s+s_i
feature set	files					
WA_*x* alone	all files	**–0.74**	–0.65	–0.70	–0.62	–0.66
best single pros. feature [6]	all files	**0.69**	0.66	0.67	0.61	0.66
pros. features on *x*	all files	**0.84**	0.78	0.79	0.72	0.79
pros. features on *x* + WA_all	all files	**0.86**	0.83	0.82	0.80	0.85
pros. features on *x* + WA_*x*	all files	—	0.78	**0.83**	0.77	0.81
pros. features on *x* + WA_all	low-error files	0.59	0.56	**0.74**	0.50	0.55
pros. features on *x* + WA_*x*	low-error files	—	0.44	**0.66**	*0.63*	*0.55*
pros. features on *x* + WA_all	high-error files	**0.86**	0.79	0.80	0.78	0.83
pros. features on *x* + WA_*x*	high-error files	—	*0.79*	0.68	0.79	**0.81**

Table 4. Regression weights for single features in the best feature sets for all 73 speech samples, depending on the scenario x and the additional use of WA_all or WA_x, respectively

scenario *x*	all	nouns	n+v	sent_i	s+s_i	nouns	n+v	sent_i	s+s_i
feature name	using WA_all					using WA_*x*			
EnNormWPW_*x* (local)	0.368	0.197	0.170	0.349	0.458	—	0.286	0.464	0.748
DurNormWPW_*x* (local)	—	—	—	—	–0.149	0.556	—	—	–0.336
F0OffW_*x* (local)	–0.198	—	–0.172	–0.141	—	—	–0.193	—	—
MeanJitter_*x* (global)	0.375	0.453	0.380	0.163	0.294	0.455	0.493	0.322	0.328
#+Voiced_*x* (global)	—	—	—	0.196	0.135	—	—	—	—
WA_all	–0.331	–0.486	–0.463	–0.445	–0.524	—	—	—	—
WA_*x*	—	—	—	—	—	—	–0.243	–0.326	–0.387
human-machine corr. *r*	0.86	0.83	0.82	0.80	0.85	0.78	0.83	0.77	0.81

12. 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.
13. 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
14. Rubenstein, H., Pickett, J.: Intelligibility of Words in Sentences. J Acoust Soc Am **30** (1958) 670
15. Smola, A., Schölkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing **14** (2004) 199–222
16. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd edn. Morgan Kaufmann, San Francisco (2005)