Offline Review Form for Submission S-ID# 9

Robust Automatic Evaluation of Intelligi...

Review by Kamil Ekštein

General instructions for the assessment: The better assessment, the higher mark, i.e. 0 = the poorest mark, 10 = the best mark (the only exception is the 'Amount of rewriting' field where 0 means 'no rewriting necessary' and 10 means 'the paper must be entirely rewritten'). Please, use your common sense or ask the organizers if unsure. This form should be filled in using Adobe Acrobat – please, do not use the built-in PDF viewer in your browser

Originality – Rate how original the work is:

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ◉ 8 ○ 9 ○ 10

Significance – Rate how significant the work is:

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Relevance – Rate how relevant the work is:

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ◉ 6 ○ 7 ○ 8 ○ 9 ○ 10

Presentation – Rate the presentation of the work:

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ◉ 8 ○ 9 ○ 10

Technical quality – Rate the technical quality of the work:

○ 0 ○ 1 ○ 2 ○ 3 ◉ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Overall rating – Rate the work as a whole:

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Amount of rewriting – Express how much of the work should be rewritten:

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Reviewer's expertise – Rate how confident you are about the above rating:

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Main contributions – Summarise main contributions:

ahoj

REVIEW ID# 1 : Robust Automatic Evaluation of Intelligi...

Positive aspects – Recapitulate the positive aspects:

Negative aspects – Recapitulate the negative aspects:

REVIEW ID# 1 : Robust Automatic Evaluation of Intelligi...

Comment (optional) – A message for the **author(s)**:

Internal comment (optional) – An internal message for the **organizers**:

After filling the form in, please, upload it to the TSD2017 web review application: Go to URL https://www.kiv.zcu.cz/tsd2017 and after logging in, please, proceed to section 'My Reviews', select the corresponding submission and press the 'Review' button. There, you'll be able to upload this PDF file.

Robust Automatic Evaluation of Intelligibility in Voice Rehabilitation Using Prosodic Analysis

(authors)

(affiliations)

Abstract. Speech intelligibility for voice rehabilitation has been successfully evaluated by automatic prosodic analysis. In this paper, the influence of reading errors and the selection of certain words for the computation of prosodic features (nouns only, nouns and verbs, beginning of each sentence, beginnings of sentences and subclauses) are examined. 73 hoarse patients (48.3 ± 16.8 years) 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. Eight prosodic features showed human-machine correlations of $r \geq 0.4$. The normalized energy in a word-pause-word interval, computed from all words ($r = 0.69$ for the full speaker set), the mean of jitter in nouns and verbs ($r = 0.67$), and the pause duration before a word ($r = 0.66$) were the most robust features. However, reading errors can significantly influence these results.

Keywords: intelligibility, automatic assessment, prosody, reading errors

1 Introduction

In speech therapy and rehabilitation, a patient's voice is usually evaluated by the therapist. Automatically computed, objective measures can support this task. However, established methods for objective evaluation, that analyze only sustained vowels, cannot evaluate speech criteria, like intelligibility. For this study, the test persons read a given standard text that underwent prosodic analysis afterwards. Earlier studies showed the suitability of this approach [3–5]. However, each prosodic feature was averaged over all words in the text and then used for further computation. Hence, content and function words, long and short words, and words at different positions in sentences, were all put together with the risk of losing information. Additionally, the influence of errors made during reading has not been analyzed in detail. When the automatic system expects the exact reproduction of a given text, then repetitions or out-of-vocabulary words have to be mapped to the pre-defined word sequence. As a consequence, the word identities and boundaries assigned by the speech recognizer are wrong. Using them for the word-based prosodic analysis leads to erroneous prosodic feature values. This problem could be solved by replacing the text reference by a transliteration of the respective speech

sample. However, this method is not applicable in clinical practice. It was shown that the influence of reading errors is neglectable for the average patient [5], but for smaller patient groups, the effects are unclear. Two main questions are addressed in this paper:

- How does the position and type of words that are selected from a read-out text influence the reliability of the automatic analysis of intelligibility?
- In what way is the automatic analysis influenced by the number of reading errors?

This work is organized as follows: Section 2 introduces the test data and the perceptual evaluation reference. The computation of the prosodic 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", [7]), 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

Table 2. Number of reading errors (in parentheses: percentual per speaker)

	all files		low-error reading		high-error reading	
	orig. files	error-treat.	orig. files	error-treat.	orig. files	error-treat.
all	226 (3.10)	149 (2.04)	11 (0.34)	9 (0.28)	215 (5.24)	140 (3.41)
substitutions	80 (1.09)	77 (1.05)	8 (0.25)	8 (0.25)	72 (1.76)	69 (1.68)
deletions	7 (0.09)	7 (0.09)	0 (0.00)	0 (0.00)	7 (0.17)	7 (0.17)
inserted words	55 (0.78)	3 (0.04)	2 (0.06)	0 (0.00)	53 (1.29)	3 (0.07)
fragments	64 (0.88)	62 (0.84)	1 (0.03)	1 (0.03)	63 (1.54)	61 (1.49)
inserted fragments	20 (0.27)	0 (0.00)	0 (0.00)	0 (0.00)	20 (0.49)	0 (0.00)

Table 3. Number of recordings with a certain number of reading errors

errors	0	1	2	3	4	5	6	7	8	11	15	17
original files	21	11	8	6	6	8	3	3	3	2	1	1
error-treated files	29	13	11	3	7	3	2	2	1	1	0	1

words and word fragments. The topic of this paper is not a full linguistic analysis of the reading errors, since the automatic analysis of intelligibility used here does not work on the linguistic level of speech. In order to describe the errors, a manual word-based counting of errors was adopted instead [5]. The three basic error classes are substitutions, deletions, and insertions. We also distinguished between substitutions of a word by another word or a word fragment, and between insertions of a full word or word fragment, respectively (Table 1, 2). We consider this word-based method sufficient since most of the errors affect one word only: the rate of single-word errors on newspaper and magazine articles among healthy speakers has been reported to be almost 70% [8].

The problem of reading errors has been addressed in two ways. In order to study the effect of errors on the evaluation on subsets of reasonable size, the overall data set was divided into an age-matched ‘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, 3). In order to determine the influence of errors within one particular data subset, a second version of the audio files was created by removing the speech parts containing additional words and fragments. Deletions, however, cannot be repaired as the correct word was not spoken in the sample. For substitutions, the situation is similar. The text flow was supposed to be preserved, so misread single words without corrections were not removed. For instance, the repetition “einst stri– einst stritten” was reduced to the correct “einst stritten” while the word “Nordwand” instead of “Nordwind” without correction was left unchanged. The data set created in this way will further be denoted as ‘error-treated’.

3 Prosodic Features

The speech recognition system used for the experiments [3] 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. This is common in automatic speech analysis on normal voices [11–13]. 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 [6]. 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 4). 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 [1, 3].

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 [1]. 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.

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:

Table 4. 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		•

– averaging over *all words* (108 words; as in earlier studies, i.e. the reference)
– *nouns only* (24 words)
– *nouns and verbs* (44 words)
– *beginnings of sentences*, i.e. the first 3 words of each of the 6 sentences (18 words)
– *beginnings of sentences and subclauses*, i.e. the first 3 words of each of the 6 sentences and 10 subclauses (48 words)

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]. 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 into shorter sections. These breaks usually occur at syntactic boundaries.

5 Results

Table 5 shows the features that for at least one of the experiments reached a human-machine correlation of $r \geq 0.4$.

The pause duration before a word (Pause–before) is only a robust indicator when it is measured before nouns. Although other scenarios, except for the beginning of sentences, also show correlations up to $r = 0.70$, the results for low-error reading are rather poor. This is supported by the correlations on error-treated files, which drop slightly when the additional utterances are removed.

The regression coefficient of the energy in a word-pause-word interval (EnRegCoeffWPW) works best when it is measured at the beginning of sentences and subunits. On the average, its human-machine correlation is $r = 0.59$, in low-error reading it decreases to $r = 0.48$; in high-error reading, $r = 0.62$ was achieved. The difference to the values on the error-treated files is not significant.

The normalized energy in a word-pause-word interval (EnNormWPW) has been reported to be a good indicator for intelligibility [3, 5]. The results in this study confirm this with $r = 0.59$ on low-error reading, $r = 0.70$ on high-error reading, and $r = 0.69$ for

Table 5. Human-machine correlation r for single local and global prosodic features ($r \geq 0.4$), depending on the words used for computation: all, nouns only, nouns and verbs (n+v), beginnings of sentences (sent_i) or of sentences and subclauses (s+s_i); bold-face: best results of each line

type	feature name	all	nouns	n+v	sent_i	s+s_i	all	nouns	n+v	sent_i	s+s_i
		all original files					all error-treated files				
local	Pause–before	0.64	0.65	**0.66**	0.35	0.51	0.62	0.64	0.65	0.32	0.47
local	EnRegCoeffWPW	0.51	0.37	0.52	0.45	**0.59**	0.48	0.31	0.49	0.46	0.58
local	EnNormWPW	**0.69**	0.64	0.59	0.59	0.66	0.68	0.62	0.59	0.59	0.65
local	DurNormWPW	0.65	**0.66**	0.63	0.43	0.56	0.64	0.65	0.62	0.43	0.55
global	MeanJitter	0.63	0.65	**0.67**	0.61	0.63	0.61	0.64	0.66	0.54	0.60
global	StandDevJitter	0.55	0.58	**0.60**	0.48	0.53	0.52	0.57	0.59	0.43	0.49
global	Dur–Voiced	0.31	0.18	0.21	0.36	0.41	0.34	0.20	0.21	0.34	**0.44**
global	DurMax–Voiced	0.36	0.21	0.24	0.32	0.42	0.38	0.21	0.23	0.30	**0.46**
		original low-error files					error-treated low-error files				
local	Pause–before	0.36	**0.62**	0.36	0.16	0.20	0.35	0.61	0.33	0.14	0.19
local	EnRegCoeffWPW	0.38	0.44	0.44	0.36	**0.48**	0.36	0.38	0.40	0.36	**0.48**
local	EnNormWPW	**0.59**	0.43	0.48	0.48	0.52	0.57	0.44	0.47	0.46	0.50
local	DurNormWPW	0.39	0.38	**0.43**	0.30	0.32	0.37	0.39	**0.43**	0.28	0.31
global	MeanJitter	0.60	0.61	**0.73**	0.57	0.57	0.60	0.61	**0.73**	0.57	0.57
global	StandDevJitter	0.50	0.53	**0.64**	0.46	0.46	0.50	0.53	0.63	0.44	0.46
global	Dur–Voiced	0.41	0.42	0.39	0.31	0.38	0.41	**0.43**	0.39	0.26	0.37
global	DurMax–Voiced	0.41	0.43	0.38	0.31	0.37	0.42	**0.44**	0.38	0.27	0.36
		original high-error files					error-treated high-error files				
local	Pause–before	0.69	0.66	0.70	0.44	0.60	0.69	0.65	**0.71**	0.46	0.57
local	EnRegCoeffWPW	0.51	0.30	0.51	0.46	**0.62**	0.48	0.24	0.49	0.46	0.61
local	EnNormWPW	**0.70**	**0.70**	0.63	0.60	0.68	**0.70**	0.67	0.63	0.61	0.67
local	DurNormWPW	0.70	**0.71**	0.66	0.50	0.62	0.70	0.70	0.65	0.52	0.62
global	MeanJitter	0.62	0.62	**0.63**	0.62	**0.63**	0.60	0.62	0.60	0.51	0.59
global	StandDevJitter	**0.58**	**0.58**	0.57	0.50	0.55	0.54	0.56	0.55	0.42	0.50
global	Dur–Voiced	0.30	0.09	0.14	0.41	0.47	0.35	0.12	0.17	0.39	**0.53**
global	DurMax–Voiced	0.36	0.13	0.21	0.36	0.47	0.40	0.14	0.21	0.34	**0.55**

the entire database, computed on the full text (Fig. 1, left). Especially for low-error reading, a selection of words from the text lowers the correlation to the perceptual scores.

The normalized duration of a word-pause-word interval (DurNormWPW) has also been a good indicator for intelligibility in earlier studies and could on the average mostly replace the energy EnNormWPW [5]. Here, it shows about the same results as the energy, but the drop for the low-error reading is much more remarkable. Only for the nouns+verbs scenario, the correlation exceeds $r = 0.40$. Both DurNormWPW and Pause–before reveal the overall speaking rate.

MeanJitter shows the highest correlation of all in this study, namely $r = 0.73$ for low-error reading and computation on nouns and verbs (Fig. 1, right). The other computation scenarios in this case are by $\Delta r \approx 0.15$ lower; for high-error reading, the correlation is

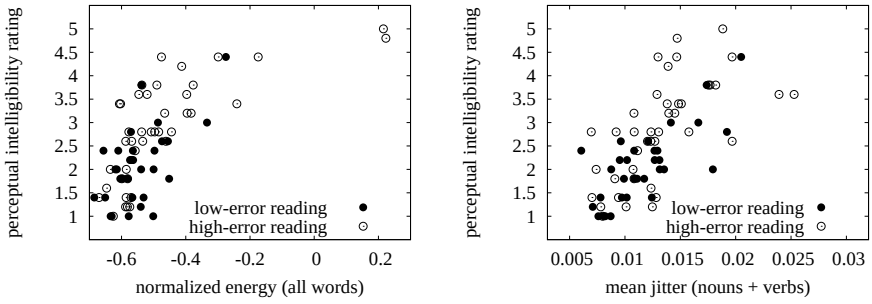

Fig. 1. Human-machine agreement: left side: for the normalized energy in a word-pause-word interval, computed on all words; right side: for the mean jitter, computed on nouns and verbs

stable at $r \approx 0.63$. In the error-treated files, only one significant drop of correlation appears when the prosodic features are computed on the beginnings of sentences.

StandDevJitter shows the same trend as MeanJitter, but with lower correlations.

The durations of the unvoiced sections in the recording (Dur–Voiced) and the longest unvoiced section (DurMax–Voiced), that contain information about the voice quality, exceed $r = 0.40$ in a few cases, but, in general, they are too unreliable to be recommended for the evaluation of intelligibility. There is a large variation among the computation scenarios: for nouns in high-error reading, only $r = 0.09$ was reached for Dur–Voiced.

We are aware of the problem arising when standard texts are used for measuring intelligibility. However, our listeners were well-trained speech therapists who were instructed to evaluate intelligibility and not voice quality. It is obvious that spontaneous speech would be the best choice for this task, and the kind of stimulus presented to the listener has an influence on the perceptual results [9]. However, spontaneous speech causes other problems. There may be a mismatch in the vocabulary of speaker and listener, the sentence structure and distribution of vowels and consonants may vary among the speakers, etc. [2]. This affects also the speech recognizer underlying the prosodic analysis. Furthermore, the prosodic evaluation of different persons is not comparable any more due to different word lengths, ratios of voiced and unvoiced sections, etc. This complexity cannot be handled properly at the moment. On the other hand, it has been shown that the text-based evaluation performed by trained listeners is as reliable as an inverse intelligibility test, where naïve raters write down a previously unknown sequence of words that was read by the test person [4].

In summary, EnNormWPW computed from all words of the text, MeanJitter of nouns and verbs, and Pause–before computed from nouns as an indicator of speaking rate, are the most robust single features for evaluation of intelligibility in this study, i.e. they show the least variability among data with different numbers of reading errors. The combination of all features and computation scenarios may reveal some more beneficial interrelations. This has been shown for features, that were averaged over the entire text, and for the average patient without regarding the reading errors. There is also room for improvement concerning the regression method, etc. This is part of future work. With

preprocessing steps of out-of-vocabulary detection and word class identification, the automatic prosodic analysis will gain even more reliability.

References

1. Batliner, A., Buckow, J., Niemann, H., Nöth, E., Warnke, V.: The Prosody Module. In: Wahlster, W. (ed.) Verbmobil: Foundations of Speech-to-Speech Translation, pp. 106–121. Springer, Berlin (2000)
2. Ellis, L., Fucci, D.: Magnitude-Estimation Scaling of Speech Intelligibility: Effects of Listeners' Experience and Semantic-Syntatic Context. Percept Mot Skills 73, 295–305 (1991)
3. (self-reference)
4. (self-reference)
5. (self-reference)
6. (self-reference)
7. International Phonetic Association (IPA): Handbook of the International Phonetic Association. Cambridge University Press, Cambridge (1999)
8. Kaufmann, R., Obler, L.: Classifiction of Normal Reading Error Types. In: Leong, C., Joshi, R. (eds.) Developmental and Acquired Dyslexia, pp. 149–157. Kluwer Academic Publishers, Dordrecht, The Netherlands (1995)
9. Kempler, D., van Lancker, D.: Effect of Speech Task on Intelligibility in Dysarthria: A Case Study of Parkinson's Disease. Brain Lang 80, 449–464 (2002)
10. Maier, A.: Speech of Children with Cleft Lip and Palate: Automatic Assessment, Studien zur Mustererkennung, vol. 29. Logos Verlag, Berlin (2009)
11. 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, 519–532 (2000)
12. 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). pp. 997–1002 (2012)
13. Rosenberg, A.: Automatic Detection and Classification of Prosodic Events. Ph.D. thesis, Columbia University, New York (2009)
14. Rubenstein, H., Pickett, J.: Intelligibility of Words in Sentences. J Acoust Soc Am 30, 670 (1958)