Intelligibility Rating with Automatic Speech Recognition, Prosodic, and Cepstral Evaluation

Tino Haderlein[1,2], Cornelia Moers[3], Bernd Möbius[4], Frank Rosanowski[2], and Elmar Nöth[1]

[1] University of Erlangen-Nuremberg, Pattern Recognition Lab (Informatik 5),
Martensstraße 3, 91058 Erlangen, Germany
Tino.Haderlein@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de

[2] University of Erlangen-Nuremberg, Department of Phoniatrics and Pedaudiology,
Bohlenplatz 21, 91054 Erlangen, Germany

[3] University of Bonn, Department of Speech and Communication,
Poppelsdorfer Allee 47, 53115 Bonn, Germany

[4] Saarland University, Department of Computational Linguistics and Phonetics,
Postfach 151150, 66041 Saarbrücken, Germany

Abstract. For voice rehabilitation, speech intelligibility is an important criterion. Automatic evaluation of intelligibility has been shown to be successful for automatic speech recognition methods combined with prosodic analysis. In this paper, this method is extended by using measures based on the Cepstral Peak Prominence (CPP). 73 hoarse patients (48.3 ± 16.8 years) uttered the vowel /e/ and read the German version of the text "The North Wind and the Sun". Their intelligibility was evaluated perceptually by 5 speech therapists and physicians according to a 5-point scale. Support Vector Regression (SVR) revealed a feature set with a human-machine correlation of $r = 0.85$ consisting of the word accuracy, smoothed CPP computed from a speech section, and three prosodic features (normalized energy of word-pause-word intervals, F_0 value at voice offset in a word, and standard deviation of jitter). The average human-human correlation was $r = 0.82$. Hence, the automatic method can be a meaningful objective support for perceptual analysis.

1 Introduction

Chronic voice diseases cause enormous costs for modern communication society [14]. A standardized, efficient method for voice assessment is therefore needed. Despite many attempts for automation, perception-based methods are still the basis for the evaluation of voice pathologies. This, however, is too inconsistent among single raters to establish a standardized and unified classification.

Perception experiments are usually applied to spontaneous speech, standard sentences, or standard texts. Automatic analysis relies mostly on sustained vowels [11]. The advantage of speech recordings is that they contain phonation onsets, variation of F_0 and pauses [13]. Furthermore, they allow to evaluate speech-related criteria, such as intelligibility. This paper focuses on automatic intelligibility assessment of chronically hoarse persons by means of automatic speech recognition, prosodic and cepstral analysis.

Most studies on automatic voice evaluation use perturbation-based parameters, such as jitter, shimmer, or the noise-to-harmonicity ratio (NHR, [11]). However, perturbation parameters have a substantial disadvantage. They require exact determination of the cycles of the fundamental frequency F_0. In severe dysphonia it is difficult to find an F_0 due to the irregularity of phonation. This drawback can be eliminated by using the Cepstral Peak Prominence (CPP) and the Smoothed Cepstral Peak Prominence (CPPS) which represent spectral noise. They do not require F_0 detection and showed high human-machine correlations in previous studies [1, 5, 8]. It is obvious that CPP expresses voice quality rather than intelligibility, but these two perceptual criteria are highly correlated with each other in voice pathologies [4]. Hence, CPP may also provide a better modeling of the perceptual concept of intelligibility.

The questions addressed in this paper are the following: How can cepstral-based evaluation support the established evaluation of intelligibility by a speech recognizer and prosodic analysis [4, 10]? Are there significant differences between the results of automatic vowel and text evaluation?

In Sect. 2, the audio data and perceptive evaluation will be introduced. Section 3 will give some information about the cepstral analysis, Sect. 4 describes the speech recognizer. An overview of the prosodic analysis and Support Vector Regression will be presented in Sect. 5 and 6, and Sect. 7 will discuss the results.

2 Test Data and Subjective Evaluation

73 German persons with chronic hoarseness (24 men and 49 women) between 19 and 85 years of age participated in this study. The average age was 48.3 years with a standard deviation of 16.8 years. Patients suffering from cancer were excluded. Each person uttered the vowel /e/ and read the text “Der Nordwind und die Sonne” (“The North Wind and the Sun”, [9]), a phonetically balanced standard text which is frequently used in medical 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 by a microphone AKG C 420.

Five experienced phoniatricians and speech scientists evaluated each speaker’s intelligibility in each recording according to a 5-point scale with the labels “very high”, “high”, “moderate”, “low”, and “none”. Each rater’s decision for each patient was converted to an integer number between 1 and 5. The average of all raters served as the reference for the automatic evaluation.

3 Cepstral Analysis

The Cepstral Peak Prominence (CPP) is the logarithmic ratio between the cepstral peak and the regression line over the entire cepstrum at this quefrency (Fig. 1). A strongly distorted voice has a flat cepstrum and a low CPP due to its inharmonic structure. The computation of CPP and the Smoothed Cepstral Peak Prominence (CPPS) was performed by the free software "cpps" [7] which implements the algorithm introduced by Hillenbrand and Houde [8]. The cepstrum was computed for each 10 ms frame, CPPS was averaged over 10 frames and 10 cepstrum bins. The vowel-based results will be denoted by "CPP-v" and "CPPS-v". For the automatic speech evaluations ("CPP-NW" and "CPPS-NW"), the first sentence only (approx. 8-12 seconds, 27 words, 44 syllables) of the read-out text was used. Sections in which the patients laughed or cleared their throat were removed from the recording.

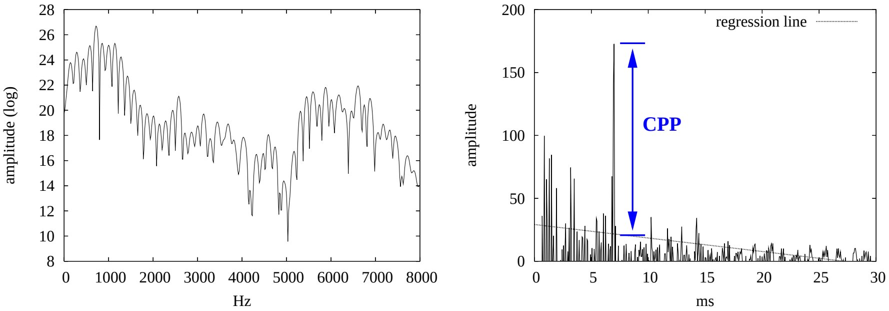

Fig. 1. Logarithmic power spectrum *(left)* and cepstrum *(right)* of a vowel section with Cepstral Peak Prominence (CPP)

4 The Speech Recognition System

The speech recognition system used for the experiments is described in detail in [17]. It is based on semi-continuous Hidden Markov Models (HMM) and can handle spontaneous speech with mid-sized vocabularies up to 10,000 words. It can model phones in any context size that is statistically useful and thus forms the so-called polyphones, a generalization of the well-known bi- or triphones. The HMMs for each polyphone have three to four states; the codebook had 500 Gaussians with full covariance matrices. The short-time analysis applies a Hamming window with a length of 16 ms, the frame rate is 10 ms. The filterbank for the Mel-spectrum consists of 25 triangle filters. For each 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 derivatives are approximated by the slope of a linear regression line over 5 consecutive frames (56 ms).

The baseline system for the experiments in this paper was trained on German dialogues of non-pathologic speakers from the Verbmobil project [18].

The data had been recorded with a close-talking microphone at a sampling frequency of 16 kHz and quantized with 16 bit. About 80% of the 578 training speakers (304 male, 274 female) were between 20 and 29 years old, less than 10% were over 40. 11,714 utterances (257,810 words) of the Verbmobil-German data (12,030 utterances, 263,633 words, 27.7 hours of speech) were used for training and 48 samples (1042 words) for the validation set, i.e. the corpus partitions were the same as in [17].

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 [4, 10]. 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 state-of-the-art in automatic speech analysis on normal voices [3, 12, 15].

The prosody module processes the output of the word recognition module and the speech signal itself. Hence, the time-alignment of the recognizer and the information about the underlying phoneme classes can be used by the module. For each speech unit of interest (here: words), a fixed reference point has to be chosen for the computation of the prosodic features. This point was chosen at the end of a word because the word is a well–defined unit in word recognition, it can be provided by any standard word recognizer, and because this point can be more easily defined than, for example, the middle of the syllable nucleus in word accent position. For each reference point, 95 prosodic features are computed from 28 base features over intervals which contain one single word, a word-pause-word interval or the pause between two words. A full description of the features used is beyond the scope of this paper; details and further references are given in [2].

In addition to the 95 local features per word, 15 global features were computed from jitter, shimmer and the number of voiced/unvoiced decisions for each

15-word interval. They cover the means and standard deviations for jitter and shimmer, the number, length and maximum length each for voiced and unvoiced sections, the ratio of the numbers of voiced and unvoiced sections, the ratio of length of voiced sections to the length of the signal and the same for unvoiced sections. The last global feature is the standard deviation of the F_0.

6 Support Vector Regression (SVR)

In order to find the best subset of word accuracy, word correctness, the prosodic features and cepstral measures to model the subjective ratings, Support Vector Regression (SVR, [16]) was used. The general idea of regression is to use the vectors of a training set to approximate a function which tries to predict the target value of a given vector of the test set. Here, the training set are the automatically computed measures, and the test set consists of the subjective intelligibility scores. For this study, the sequential minimal optimization algorithm (SMO, [16]) of the Weka toolbox [19] was applied in a 10-fold cross-validation manner.

The pre-defined correlation-based feature selection algorithm [6] had been altered so that the number of matrix inversions was substantially reduced at the cost of a slightly worse selection result [10, pp. 59-61]. The features with the highest ranks were used as input for the SVR.

7 Results and Discussion

The correlations between the perceptual evaluation and single automatic measures are given in Table 1. The human-machine correlations of these measures alone are not as good as the inter-rater correlation of a panel of experts (Table 2). But it appears that WA outperforms WR, and the text-based cepstral measures are clearly better than the vowel-based ones. The correlations are negative because high recognition rates and cepstral peaks came from "good" voices with a low score and vice versa. The values did not change significantly throughout the study when Spearman's rank-order correlation ρ was computed. For this reason, only Pearson's r is given.

By using WA, WR, the CPP measures, and the prosodic features as input for SVR, higher correlations to the subjective intelligibility score were obtained (Table 3). The WR and the vowel-based CPP measures did not appear in the selected feature list. A human-machine correlation of $r = 0.85$ was achieved for the set of WA, CPPS-NW, the normalized energy of word-pause-word intervals (EnNormWPW), the F_0 value at the voice offset in a word (F0OffWord), and the standard deviation of jitter (StandDevJitter). With the latter three prosodic features alone, $r = 0.79$ was measured. CPPS-NW and WA together reach $r = 0.83$. The other selected experiments given in Table 3 show that for a human-machine correlation of $r \geq 0.80$ either WA or CPPS-NW are needed in any case.

The energy value EnNormWPW is normalized with respect to a pre-computed speaker list. If the person has a hoarse and irregular voice, then the energy level especially in the high frequency portions is raised. For this reason, this feature may contribute strongly to the best feature set. The impact of the F_0 value can be explained by the noisy speech that causes octave errors during F_0 detection, i.e. instead of the real fundamental frequency, one of its harmonics is found. With more "noisy speech", this may influence the F_0 trajectory and hence the correlation to the subjective results. It is not clear so far, however, why only the end of the voiced sections causes a noticeable effect. There may be a connection to changes in the airstream between the beginning and the end of words or phrases. It may have its reason in the high speaking effort which leads to more irregularities especially in these positions, but this has to be confirmed by more detailed experiments. Jitter is one of the established measures for voice pathology. However, a certain amount of jitter and regular changes thereof are present in normal voices. When changes of jitter over time become irregular, this may also be an indicator for a less intelligible voice. Note that the prosody module computes the F_0 and jitter values only on sections which it has previously identified as voiced.

The correlations between the feature values of the best subset are given in Table 4. A high EnNormWPW correlates significantly with a low CPPS-NW and a low WA. Likewise, jitter and its standard deviation are higher which correlates negatively with CPPS-NW. The low CPPS-NW in a distorted voice correlates with a low recognition rate.

Table 1. Subjective and objective evaluation results for 73 hoarse speakers: intelligibility, word accuracy (WA) and word correctness (WR), and the cepstral peak measures; the rightmost column shows the correlation r between the human evaluation and the respective automatic measure

measure	unit	mean	st. dev.	min.	max.	r
intell.	points	2.5	1.0	1.0	5.0	*1.00*
WA	%	69.3	14.3	27.8	90.1	–0.74
WR	%	73.5	12.0	28.9	90.1	–0.69
CPP-v	dB	17.2	4.3	8.8	25.3	–0.61
CPPS-v	dB	6.1	2.2	0.9	11.1	–0.58
CPP-NW	dB	12.1	1.6	9.1	16.3	–0.69
CPPS-NW	dB	4.1	1.0	1.9	6.3	–0.74

Table 2. Inter-rater correlation r for intelligibility between each rater and the average of the remaining raters

rater	K	R	S	T	V	avg.
r	0.78	0.84	0.88	0.75	0.84	0.82

Table 3. SVR regression weights for the best subset (experiment 1) and selected other subsets, and their correlation r to the subjective intelligibility scores (last row)

feature	exp. 1	exp. 2	exp. 3	exp. 4	exp. 5	exp. 6	exp. 7
EnNormWPW	0.228	0.980	0.840		0.345	0.660	
F0OffWord	–0.146	–0.428					
StandDevJitter	0.167	0.522	0.549	0.168	0.343	0.178	
CPPS-NW	–0.412			–0.485		–0.524	–0.632
WA	–0.431			–0.579	–0.532		–0.539
correlation r	0.85	0.79	0.74	0.84	0.83	0.81	0.83

Table 4. Correlation of the feature values of the best feature set for all 73 speakers

feature	F0OffWord	StandDevJitter	CPPS-NW	WA
EnNormWPW	–0.03	0.23	–0.56	–0.74
F0OffWord		–0.10	0.26	0.09
StandDevJitter			–0.58	–0.30
CPPS-NW				0.56

For this study, patients read a standard text, and voice professionals evaluated intelligibility. It is often argued that intelligibility should be evaluated by an "inverse intelligibility test": The patient utters a subset of words and sentences from a carefully built corpus. A naïve listener writes down what he or she heard. The percentage of correctly understood words is a measure for the intelligibility of the patient. However, when automatic speech evaluation is performed for instance with respect to prosodic phenomena, such as word durations or percentage of voiced segments, then comparable results for all patients can only be achieved when all the patients read the same defined words or text. This means that an inverse intelligibility test can no longer be performed, and intelligibility has to be rated on a grading scale instead.

The results obtained in this study allow for the following conclusions: There is a significant correlation between subjective rating of intelligibility and automatic evaluation. The human-machine correlation is better than the average inter-rater correlation among speech experts. Cepstral-based measures improve the human-machine correlation, but only when they are computed from a speech recording and not from a sustained vowel only. The method can serve as the basis for an automatic, objective system that can support voice rehabilitation.

Acknowledgments

This work was partially funded by the German Cancer Aid (Deutsche Krebshilfe) under grant 107873. The responsibility for the contents of this study lies with the authors. We would like to thank Dr. Hikmet Toy for acquiring and documenting the audio data.

References

1. Awan, S., Roy, N.: Outcomes Measurement in Voice Disorders: Application of an Acoustic Index of Dysphonia Severity. J. Speech Lang. Hear. Res. 52, 482–499 (2009)
2. Batliner, A., Buckow, J., Niemann, H., Nöth, E., Warnke, V.: The Prosody Module. In: Wahlster [18], pp. 106–121
3. Chen, K., Hasegawa-Johnson, M., Cohen, A., Borys, S., Kim, S.S., Cole, J., Choi, J.Y.: Prosody dependent speech recognition on radio news corpus of American English. IEEE Trans. Audio, Speech, and Language Processing 14, 232–245 (2006)
4. Haderlein, T.: Automatic Evaluation of Tracheoesophageal Substitute Voices, Studien zur Mustererkennung, vol. 25. Logos Verlag, Berlin (2007)
5. Halberstam, B.: Acoustic and Perceptual Parameters Relating to Connected Speech Are More Reliable Measures of Hoarseness than Parameters Relating to Sustained Vowels. ORL J. Otorhinolaryngol. Relat. Spec. 66, 70–73 (2004)
6. Hall, M.: Correlation-based Feature Subset Selection for Machine Learning. Ph.D. thesis, University of Waikato, Hamilton, New Zealand (1999)
7. Hillenbrand, J.: cpps.exe [software], available at: http://homepages.wmich.edu/˜hillenbr. Accessed May 30, 2011
8. Hillenbrand, J., Houde, R.: Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech. J. Speech Hear. Res. 39, 311–321 (1996)
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, Studien zur Mustererkennung, vol. 29. Logos Verlag, Berlin (2009)
11. Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., Corthals, P.: Acoustic measurement of overall voice quality: A meta-analysis. J. Acoust. Soc. Am. 126, 2619–2634 (2009)
12. 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)
13. Parsa, V., Jamieson, D.: Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. J. Speech Lang. Hear. Res. 44, 327–339 (2001)
14. Ruben, R.: Redefining the survival of the fittest: communication disorders in the 21st century. Laryngoscope 110, 241–245 (2000)
15. Shriberg, E., Stolcke, A.: Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing. In: Proc. International Conference on Speech Prosody. pp. 575–582. Nara, Japan (2004)
16. Smola, A., Schölkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing 14, 199–222 (2004)
17. Stemmer, G.: Modeling Variability in Speech Recognition, Studien zur Mustererkennung, vol. 19. Logos Verlag, Berlin (2005)
18. Wahlster, W. (ed.): Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin (2000)
19. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition. Morgan Kaufmann, San Francisco (2005)