Automatic Rating of Hoarseness by Text-based Cepstral and Prosodic Evaluation

Tino Haderlein[1,2], Cornelia Moers[3], Bernd Möbius[4], 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 (now with Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands)
[4] Saarland University, Department of Computational Linguistics and Phonetics
Postfach 151150, 66041 Saarbrücken, Germany

Abstract. The standard for the analysis of distorted voices is perceptual rating of read-out texts or spontaneous speech. Automatic voice evaluation, however, is usually done on stable sections of sustained vowels. In this paper, text-based and established vowel-based analysis are compared with respect to their ability to measure hoarseness and its subclasses. 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". Five speech therapists and physicians rated roughness, breathiness, and hoarseness according to the German RBH evaluation scheme. The best human-machine correlations were obtained for measures based on the Cepstral Peak Prominence (CPP; up to $|r| = 0.73$). Support Vector Regression (SVR) on CPP-based measures and prosodic features improved the results further to $r \approx 0.8$ and confirmed that automatic voice evaluation should be performed on a text recording.

1 Introduction

Evaluation of voice distortions is still mostly performed perception-based. Perception of voice qualities, however, is too inconsistent among single raters to establish a standardized and unified classification. The average opinion of a panel of raters is more consistent, but this approach is not suitable for clinical application. The ideal solution would be objective, automatic assessment.

The perception experiments are applied to spontaneous speech, read-out standard sentences, or standard texts. In contrast, already used methods of automatic analysis rely mostly on sustained vowels [13]. The advantage of speech recordings, however, is that they contain phonation onsets, variation of F_0 and pauses [16]. Furthermore, they allow to evaluate speech-related criteria, such as intelligibility [5]. For this reason, also for the automatic evaluation, speech recordings should be used. This paper focuses on

P. Sojka et al. (Eds.): TSD 2012, LNCS 7499, pp. 573–580, 2012.

the automatic assessment of hoarseness and its subclasses by means of prosodic and cepstral analysis.

Hoarseness is a psycho-acoustically defined measure which was originally believed to be distinct of the other two categories roughness (or harshness) and breathiness. Nowadays, hoarseness is often seen as the superclass of these categories [1]. The Roughness-Breathiness-Hoarseness (RBH) evaluation scheme [15] takes this into account. It is an established means for perceptual voice assessment in German-speaking countries and serves as the reference for the automatic analysis presented in this paper.

Most studies on automatic voice evaluation use perturbation-based parameters, such as jitter, shimmer, or the noise-to-harmonicity ratio (NHR, [13]). 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 [2,6,9].

The questions addressed in this paper are the following: How do cepstral- and text-based prosodic measures perform in comparison to established, vowel-based measures? How well does cepstral-based analysis correspond to perception-based RBH evaluation when it is supported by prosodic analysis?

In Sect. 2, the audio data and perceptual evaluation will be introduced. Section 3 will give some information about the cepstral analysis, Sect. 4 describes the vowel analysis with the Praat software. 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 subjects 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”, [11]), 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 using an AKG C 420 microphone (AKG Acoustics, Vienna, Austria).

The text recordings were evaluated perceptually by 5 speech therapists and physicians according to the German Roughness-Breathiness-Hoarseness (RBH) scale [15]. Each of the three criteria can be evaluated on a 4-point scale where ‘0’ means “absent” and ‘3’ means “high degree”. In order to capture the fact that hoarseness is the superclass, the H rating must have either the same or a higher rating than R or B. RBH represents a short version of the GRBAS scale [10], with the categories “asthenia” and “strain” omitted.

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. 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 means of the free software "cpps" [8] which implements the algorithm introduced by Hillenbrand and Houde [9]. 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 for this pilot experiment by hand.

4 Analysis of Sustained Vowels with Praat

The automatic analysis of the sustained vowels (/e/) with respect to established irregularity measures was performed using the software Praat 5.1 [4]. An overview of the features is given in Table 1. For this vowel analysis, sections of at least 0.5 seconds duration of stable phonation excluding onset and offset were evaluated. From 17 speakers, a section of 0.7 seconds could be extracted; from 36 speakers, a full second was available. Although several measures are F_0-based, men and women were not analyzed separately for this study. The reason is that the goal of the analysis was to find measures which can be used independently of the speaker's gender, just like the human raters do not need different evaluation methods for men and women.

5 Prosodic Features

In order to find automatically computable counterparts for the RBH criteria, also a "prosody module" was used to compute features based upon frequency, duration and speech energy (intensity) measures.

The prosody module processes the output of a word recognition module [5] and the speech signal itself. Hence, it can use the time-alignment of the recognizer and the information about the underlying phoneme classes. 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, 28 prosodic features are computed which refer to a single word or the pause between two words. Ten of these features are additionally computed for a word-pause-word interval. A full description of the features used is beyond the scope of this paper; details and further references are given in [3].

In addition to the 38 local features per word position, 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 and ratio of the durations 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 F_0.

6 Support Vector Regression (SVR)

In order to find the best subset of the prosodic features and cepstral measures to model the subjective ratings, Support Vector Regression (SVR, [17]) 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 comprises the automatically computed measures, and the test set consists of the subjective RBH scores. For this study, the sequential minimal optimization algorithm (SMO, [17]) of the Weka toolbox [18] was applied in a 10-fold cross-validation manner.

7 Results and Discussion

For all three perceptual RBH criteria, the entire range between 0 and 3 was covered. The average values were 1.56 (standard deviation: 0.83) for R, 1.19 (0.81) for B, and 1.84 (0.84) for H (see also [14]).

The correlations between the perceptual evaluation and the single automatic measures are given in Table 1. For all evaluated criteria, the best results for single objective measures were obtained for CPPS-NW. The text-based CPP-NW and CPPS-NW perform remarkably better than all the vowel-based measures, including CPP-v and CPPS-v.

The SVR on all available measures, i.e. the Praat and CPP-based features, and the prosodic features, revealed higher correlations than on the single measures. Two feature sets were identified, where one was optimal for R and H and the other one for B (Table 2). For all criteria, CPPS-NW, the normalized energy in word-pause-word intervals (EnNormWPW), the average minimal F_0 of each word (F0MinWord), and the average F_0 at voice offset (F0OffWord) were in the respective set. Additionally, the duration of all unvoiced sections in the signal (Dur–Voiced) was part of the best set for breathiness. None of the vowel-based measures contributed to the best feature sets. This clearly indicates the need for text-based automatic evaluation.

The energy value EnNormWPW may contribute strongly to the best feature sets, because it was normalized with respect to healthy speakers [3]. Loudness effects are removed by the normalization, but if a person has a hoarse and irregular voice, then the energy level especially in the high frequency portions is higher than for normal speakers. The impact of the F_0 values F0MinWord and F0OffWord 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,

Table 1. Correlation *r* between perceptual and automatic evaluation (**: significant on the 0.01 level, *: significant on the 0.05 level); the perceptual result was the mean value of all raters. The names of the features in the first part of the table follow the names used in Praat [4]; APQ11 was computed for 72 patients only due to an invalid value for one patient.

data	feature(s)	R	B	H
vowel	Jitter local	0.33**	0.54**	0.51**
vowel	Jitt local absolute	0.33**	0.28*	0.34**
vowel	RAP (Rel. Avg. Perturb. Quotient), jitter of 3 periods	0.26*	0.39**	0.38**
vowel	PPQ5 (Pitch Perturb. Quotient), jitter of 5 periods	0.24*	0.32**	0.33**
vowel	Shimmer local	0.38**	0.56**	0.58**
vowel	Shimmer local absolute	0.39**	0.56**	0.59**
vowel	APQ11 (Amplitude Perturb. Quotient of 11 periods)	0.34**	0.41**	0.47**
vowel	NHR (N of [1500;4500] Hz / H of [70;4500] Hz)	0.34**	0.54**	0.53**
vowel	HNR (Mean harmonicity-to-noise ratio)	–0.40**	–0.57**	–0.59**
vowel	CPP-v (Cepstral Peak Prominence, vowel-based)	–0.25*	–0.60**	–0.53**
vowel	CPPS-v (smoothed CPP, vowel-based)	–0.17	–0.52**	–0.44**
text	CPP-NW (Cepstral Peak Prominence, text-based)	–0.47**	–0.69**	–0.69**
text	CPPS-NW (smoothed CPP, text-based)	–0.52**	–0.69**	–0.73**
v.+t.	SVR on all features: best set for B	0.72**	**0.82****	0.77**
v.+t.	SVR on all features: best set for R and H	**0.74****	0.79**	**0.79****

Table 2. SVR regression weights for the best feature subsets when predicting the perceptual RBH scores

	best set for R, H			best set for B		
predicted score	R	B	H	R	B	H
EnNormWPW	–0.063	0.565	0.372	0.014	0.589	0.389
F0MinWord	–0.678	0.076	–0.431	–0.562	–0.562	–0.408
F0OffWord	–0.223	–0.410	–0.274	–0.261	–0.261	–0.269
Dur–Voiced	———			0.470	0.262	0.044
CPPS-NW	–0.444	–0.612	–0.615	–0.215	0.641	–0.631

Table 3. Correlation *r* of the feature values of the best feature sets

feature	F0MinWord	F0OffWord	Dur–Voiced	CPPS-NW
EnNormWPW	0.12	–0.03	0.09	–0.56
F0MinWord		0.11	–0.56	0.37
F0OffWord			–0.04	0.26
Dur–Voiced				–0.49

however, why in the case of F0OffWord only the end of the voiced sections causes a noticeable effect. It may reflect changes in the airstream between the beginning and the end of words or phrases. High speaking effort leads to more irregularities especially in these positions, but this has to be confirmed by more detailed experiments. Note that the prosody module computes the F_0 values only on sections which it has previously identified as voiced. It may be this property of the software that makes F_0-based values so important for the analysis of distorted voices, although the purpose of adding cepstral measures to the feature set was to become independent of them.

The duration of the voiceless sections in the signal (Dur–Voiced) is comparable among all speakers since they read the same text. Hence, a higher duration indicates a higher percentage of voiceless sections and thus an irregular voice.

The correlations between the feature values of the best subsets are given in Table 3. A low EnNormWPW correlates significantly with a high CPPS-NW, because both indicate a high-quality voice. Likewise, CPPS-NW and Dur–Voiced are negatively correlated. A large duration of unvoiced sections correlates negatively with the average F_0 minimum. The reason may be the predefined F_0 threshold of the prosody module. The lowest F_0 that will be returned is 50 Hz, lower values are classified as voiceless. Hence, a very low voice will result in a low minimal F_0 of about 50 Hz and in a higher amount of unvoiced sections caused by F_0 values below the threshold.

The average inter-rater correlation between one rater and the average of the other ones was $r = 0.76$ for R, $r = 0.70$ for B, and $r = 0.82$ for H. For breathiness, the text-based CPP values ($r = 0.69$) alone almost reached the human reference, the SVR results even outperformed it. For roughness and hoarseness, the SVR almost reached the human inter-rater correlation.

The results on human-machine correlation with cepstral parameters confirmed some findings of other studies. Hillenbrand and Houde [9] found a significant correlation between these parameters and the perceived degree of breathiness for sustained vowels and speech recordings. This was confirmed in our study, but only for speech recordings. Heman-Ackah et al. [7] reported a correlation of the total degree of dysphonia and CPPS of $r = –0.80$ on stable vowel sections and $r = –0.86$ on sentence recordings. Their correlation between vowel- and sentence-based CPPS and the breathiness rating was $r = –0.70$ and $r = –0.71$, respectively. Our best result for a single feature for breathiness was $r = –0.69$ for both CPP-NW and CPPS-NW. The vowel-based measures, however, reached just $r = –0.60$. Nevertheless, breathiness was better modeled by the vowel-based CPP-v than roughness or hoarseness. The most probable reason for the differences in vowel analysis among the studies is that it requires stable phonation. Often a frame of one second of the vowels /a/ (predominantly), /e/, or /i/ is chosen. Other vowel segment durations from 0.1 seconds up to 3 seconds have been reported [13]. For our study, the minimum duration of stable phonation was set to 0.5 seconds, because some patients were not able to phonate longer without too much irregularity. Our subjects uttered /e/, sometimes shifted towards /ɛ/ which is the adjacent phoneme in the German vowel space. Therefore, the results may not be completely comparable to other studies. On the other hand, these variations in duration and vowel quality show that there will always be inconsistencies in the data obtained from a representative group of patients. If their influence deteriorates the evaluation results that much, then the method cannot

be used for clinical purposes. This is another important argument against vowel-based perturbation analysis for voice evaluation.

CPP and CPPS cannot differentiate between different voice qualities [2]. Even with support by prosodic features, this is not possible with the available feature set, because the best sets for modeling R, B, and H are too similar. When these sets were applied to predict all of the rating criteria (Table 2), still a few clear differences in the weighting factors for the prediction formulae were revealed. The normalized energy EnNormWPW, for instance, could also be left out of the set for R. The results change only marginally then. This feature is obviously more important for the evaluation of breathiness and overall hoarseness.

The similarity of the best feature sets for all rating criteria is consistent with the interaction between different dimensions of human perception: the presence of roughness in a voice does not influence the perception of breathiness. However, the perceived degree of roughness is strongly influenced by the presence of breathiness [12]. Additionally, dysphonic voices with lower fundamental frequencies are perceived as being more rough than those with higher F_0 [19]. Our results, however, confirm the assumption that roughness and breathiness are perceived as separate dimensions and that hoarseness is the superclass of both [1]. R and H showed a better correlation with each other in the human results ($r = 0.81$) than B and H ($r = 0.76$). The correlation of R and B was only $r = 0.36$. It is remarkable that breathiness and hoarseness were better mapped by the automatically obtained measures than roughness, even more so since the optimal feature sets for R and H were the same. It will be one of the most important aspects in future work to teach the automatic analysis to tell apart roughness, breathiness, and hoarseness as well as human listeners do.

Some aspects for enhancing the human-machine correlations have not been tested in this study. Human perception is often non-linear, such as indicated by the Bark scale for pitch, for instance. Physical scales are often linear, such as the frequency measured in Hertz. Better human-machine correlations may be found with non-linear mappings between the two modalities.

8 Conclusion

The results obtained in this study allow for the following conclusions: There is a significant correlation between the subjective rating of roughness, breathiness, hoarseness, on the one hand, and the automatic evaluation, on the other. However, the three criteria cannot be rated separately with the available set of features. The human-machine correlation is about as good as 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 Else Kröner-Fresenius-Stiftung (Bad Homburg v.d.H., Germany) under grant 2011_A167. 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. Aronson, A., Bless, D.: Clinical Voice Disorders. Thieme, 4th edn. (2009)
2. 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)
3. 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)
4. Boersma, P., Weenink, D.: Praat: Doing phonetics by Computer, Version 5.1.33, `http://www.fon.hum.uva.nl/praat` (accessed May 21, 2012)
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.) TSD 2011. LNCS, vol. 6836, pp. 195–202. Springer, Heidelberg (2011)
6. 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)
7. Heman-Ackah, Y., Michael, D., Goding Jr., G.: The Relationship Between Cepstral Peak Prominence and Selected Parameters of Dysphonia. J. Voice 16, 20–27 (2002)
8. Hillenbrand, J.: cpps.exe (software), `http://homepages.wmich.edu/~hillenbr` (accessed May 21, 2012)
9. Hillenbrand, J., Houde, R.: Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech. J. Speech Hear. Res. 39, 311–321 (1996)
10. Hirano, M.: Clinical Examination of Voice. Springer, New York (1981)
11. International Phonetic Association (IPA): Handbook of the International Phonetic Association. Cambridge University Press, Cambridge (1999)
12. Kreiman, J., Gerratt, B., Berke, G.: The multidimensional nature of pathologic vocal quality. J. Acoust. Soc. Am. 96, 1291–1302 (1994)
13. 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)
14. Moers, C., Möbius, B., Rosanowski, F., Nöth, E., Eysholdt, U., Haderlein, T.: Vowel- and Text-based Cepstral Analysis of Chronic Hoarseness. J. Voice 26, 416–424 (2012)
15. Nawka, T., Anders, L.C., Wendler, J.: Die auditive Beurteilung heiserer Stimmen nach dem RBH-System. Sprache - Stimme - Gehör 18, 130–133 (1994)
16. Parsa, V., Jamieson, D.: Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. J. Speech Lang. Hear. Res. 44, 327–339 (2001)
17. Smola, A., Schölkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing 14, 199–222 (2004)
18. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
19. Wolfe, V., Martin, D.: Acoustic Correlates of Dysphonia: Type and Severity. J. Commun. Disord. 30, 403–416 (1997)