IEEE copyright notice

ROBUST ESTIMATION OF VOICE QUALITY PARAMETERS UNDER REAL WORLD DISTURBANCES

M. Lugger and B. Yang

Chair of System Theory and Signal Processing
University of Stuttgart, Germany
Marko.Lugger@lss.uni-stuttgart.de

W. Wokurek

Institute of Natural Language Processing
University of Stuttgart, Germany
wokurek@ims.uni-stuttgart.de

ABSTRACT

This paper investigates the influence of different types of disturbances to the estimation of voice quality parameters. Here, voice quality is not only based on the pitch or pitch contour as in many approaches. The parameters are estimated by spectral gradients of the vocal tract compensated speech signal. We present a set of five parameters for describing the voice quality. They are used to distinguish between gender, voice qualities, and many emotional states of the speaker. We estimate them from speech signals which are corrupted by background noise and room reverberation. The paper demonstrates a certain degree of robustness of the voice quality parameters against these real world disturbances.

1. INTRODUCTION

Voice quality, in contrast to other areas in speech processing, is a less explored field. It is a paralinguistic i.e. nonverbal part of speech communication. Voice quality describes the kind of phonation of speech utterances like modal, breathy or creaky voice. Emotions may influence voice quality. Together with pitch and duration, voice quality contributes to speech prosody. By evaluating the voice quality, the listener obtains information about the physical, psychological, and emotional characteristics of a speaker.

Former investigations have shown that the voice quality parameters allow the distinction between different speaking groups like gender [1], pathological and nonpathological speakers [2], and word stress [3]. Other potential applications in speech analysis are emotion detection, speaker identification, and improved speech recognition. One application in speech synthesis is the improved naturalness of the synthesized speech by incorporating voice quality features.

Unfortunately, all studies up to now assumed clean speech signals. In this paper, we study the robustness of voice quality parameters under background noise and room reverberations. The voice quality parameters are tested in three applications concerning gender, voice quality, and emotion.

The paper is organized as follows. Section 2 presents the voice quality estimators. In section 3, the speech data of different applications and their disturbances are described. The robustness of the voice quality parameters is studied detailed in section 4.

This work was supported by a grant from the Ministry of Science, Research and the Arts of Baden-Württemberg, Germany

2. VOICE QUALITY

Voice quality is mainly affected by the excitation of the human voice that is called phonation. This means that the shape of the glottal pulse is responsible for the voice quality that a speaker is realizing. In contrast, all procedures that belong to the articulation process affect the articulated sounds that all together build the linguistic content of the speech.

Usually voice quality parameters are obtained from the electroglottographical signal which is measured directly at the glottis. Electroglottography (EGG) is a technique used to record the laryngeal behavior indirectly by measuring the change in electrical conductivity across the throat during speaking.

2.1. Voice Quality Parameters

In the literature there are other methods for estimating parameters closely related to the voice quality or emotion. For voice quality, the most common method is fitting a glottal pulse model to the inverse filtered speech signal [4]. For detecting emotions many approaches are mainly based on the pitch contour of the speaker [5]. Some works involve prosodic characteristics like intensity, word stress or rate of speech [6].

We propose a method to estimate the voice quality parameters directly from the acoustic speech signal. No extra hardware and no invasion to the human body are required to obtain the desired information. The method is based on the observations by Stevens and Hanson [7] that the glottal properties ”open quotient”, ”glottal opening”, ”skewness of glottal pulse”, and ”rate of glottal closure” each affect the excitation spectrum of the speech signal in a dedicated frequency range and thus reflect the voice quality of the speaker. They proposed to estimate these glottal states from the acoustic speech signal by adequate relation of the amplitudes of the corresponding higher harmonics with that of the fundamental mode. They further found that the first formant bandwidth is correlated with the incompleteness of the glottal closure. These measurements are simply called voice quality parameters.

Our modified algorithm calculates spectral gradients instead of pure amplitude ratios, because gradients better char-

acterize the shape of the glottal signal spectrum. In addition, a vocal tract compensation is performed prior estimating the gradients [1]. The whole process can be divided into three steps: measurement of speech features, compensation of the vocal tract influence, and estimation of the voice quality parameters. Below we describe these steps in more details.

2.2. Measurement of speech features

The first step estimates some well known speech features from windowed, voiced segments of the speech signal. We perform the voiced-unvoiced decision and the pitch estimation according to the RAPT algorithm [8] that looks for peaks in the normalized cross correlation function. The frequencies and bandwidths of the first four formants are estimated by an LPC analysis [9]. All frequency values are converted to the Bark scala.

Feature	Meaning
F_p	pitch
F_1, F_2, F_3, F_4	formant frequencies
B_1, B_2, B_3, B_4	formant bandwidths
H_1, H_2	amplitude at F_p and $2F_p$
$F_{1p}, \ldots, F_{3p}$	frequency of spectrum peaks near formants
$A_{1p}, \ldots, A_{3p}$	amplitude values at $F_{1p}, \ldots, F_{3p}$

Table 1. Speech features used for voice quality parameter estimation

2.3. Compensation of the vocal tract influence

Since the voice quality parameters shall only depend on the excitation and not on the articulation process, the influence of the vocal tract has to be compensated. The contribution of each of the four formants to the spectrum at frequency f is estimated by [10]

$$V(f; F_i, B_i) = 20 \log \frac{F_i^2 + \left(\frac{B_i}{2}\right)^2}{\sqrt{\left((f-F_i)^2 + \left(\frac{B_i}{2}\right)^2\right)\left((f+F_i)^2 + \left(\frac{B_i}{2}\right)^2\right)}}$$

They are removed from the amplitudes H_k and A_{kp} in Table 1:

$$\tilde{H}_k = H_k - \sum_{i=1}^{4} V(kF_p; F_i, B_i) \qquad (k = 1, 2)$$

$$\tilde{A}_{kp} = A_{kp} - \sum_{\substack{i=1 \\ i \neq k}}^{4} V(F_{kp}; F_i, B_i) \qquad (k = 1, 2, 3)$$

The results of the formant compensation are the corrected spectral amplitudes $\tilde{H}_1, \tilde{H}_2$ of the first and second harmonics and the corrected peak amplitudes $\tilde{A}_{1p}, \tilde{A}_{2p}$, and $\tilde{A}_{3p}$ near the three formants as shown in Figure 1.

2.4. Estimation of the voice quality parameters

The last step estimates the following five voice quality parameters from the vocal tract compensated speech features:

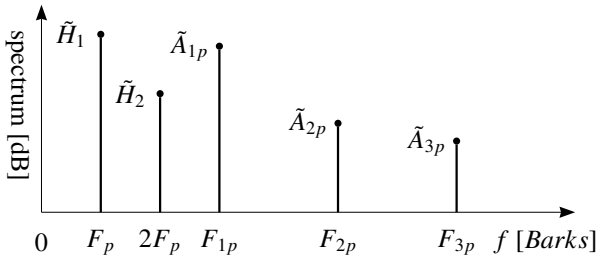

Fig. 1. Vocal tract compensated peaks of the FFT spectrum for the voice qulity parameter estimation

"**O**pen **Q**uotient **G**radient", "**G**lottal **O**pening **G**radient", "**SK**ewness **G**radient", "**R**ate of **C**losure **G**radient", and "**I**ncompleteness of **C**losure". They are given by

$$\begin{aligned} \text{OQG} &= \frac{\tilde{H}_1 - \tilde{H}_2}{F_p} \\ \text{GOG} &= \frac{\tilde{H}_1 - \tilde{A}_{1p}}{F_{1p} - F_p} \\ \text{SKG} &= \frac{\tilde{H}_1 - \tilde{A}_{2p}}{F_{2p} - F_p} \\ \text{RCG} &= \frac{\tilde{H}_1 - \tilde{A}_{3p}}{F_{3p} - F_p} \\ \text{IC} &= \frac{B_1}{F_1} \end{aligned}$$

Figure 2 gives an illustration of the first four parameters as spectral gradients with respect to the pitch frequency.

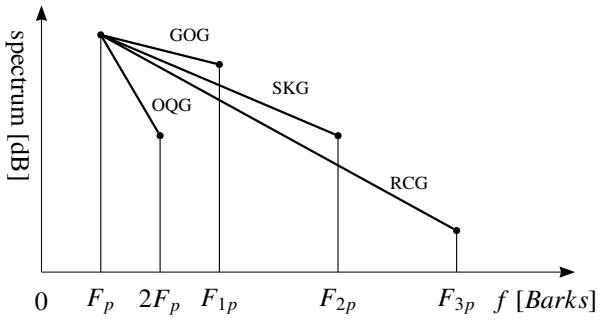

Fig. 2. Voice quality parameters as spectral gradients

3. EXPERIMENTS: SPEECH AND DISTURBANCES

We use the voice quality parameters to distinguish between different speaking groups in three studies: gender, voice qualities according to Laver, and emotional states. All speech data are recorded in an anechoic room at the sampling frequency of 16 kHz.

3.1. Distinction between gender

The speech samples for the first study ware taken from [11]. They consist of 10 utterances of male and 10 utterances of female speech, each of about 30 second duration.

3.2. Distinction between voice qualities

The speech data for the second study are taken from the book [12] by Laver. It contains utterances spoken in six different voice qualities [13] by the same speaker: "modal voice", "falsetto voice", "whispery voice", "breathy voice", "creaky voice", and "rough voice".

3.3. Distinction between emotional states

The speech data for the third study are from the "Berlin database of emotional speech" [14]. It contains about 500 sentences spoken by actors in a neutral, happy, angry, sad, fearful, bored, and disgusted way.

3.4. Real world disturbances

In order to study the robustness of voice quality parameters under real world disturbances, the speech signals mentioned above are distorted by noise and reverberation prior to estimating the voice quality parameters.

3.4.1. Noise

Acoustic background noise can be very versatile in its characteristics. It reaches from white noise over coloured noise to time varying noise and the so called cocktail party noise that consists of dozens of people talking in the background. We add the following noise to the speech signal: "white noise", "pink noise", "factory noise" and "cocktail party noise". The degree of the noise is controlled by the global signal-to-noise-ratio (SNR)

$$SNR(N) = 10 \log \frac{\sum_{k=1}^{N} s^2(k)}{\sum_{k=1}^{N} n^2(k)} \quad [dB]$$

where $s(k)$ is the speech signal, $n(k)$ the noise signal and N the total number of speech samples.

3.4.2. Room Reverberation

Room acoustic effects can be modelled by the room impulse response

$$h(t) = \sum_{n} a_n \delta(t - t_n)$$

Every tap a_n represents one path from the source to the microphone. One measure for the degree of room acoustic is the reverberation time T_{60}. It is defined as the time in seconds for the reverberation level to decay to 60 dB below the initial level. In our experiments, simulated room impulse responses according to the image method [15] are used.

4. RESULTS

The distinction between different speaking groups by using voice quality parameters is realized by t-tests. The significance level for all analysis was set to $\alpha = 0.05$. In the Tables 2-4, the observed one-sided level of significance for gender distinction is shown. A value smaller (larger) than α indicates that the male and female speech signals can (not) be distinguished from each other. Values smaller than α in in boldface.

Table 2 shows the test results under white noise for a varying SNR. For clean speech with SNR $=\infty$, all voice quality parameters allow a distinction. If we decrease SNR to -10 dB, RCG and IC lose their discrimination capabilities while OQG, GOG, and SKG are robust and stay significant. The noise of a factory hall in Table 3 shows a similar behavior. Only the parameters OQG and GOG become unsignificant for lower SNR values. Table 4 shows the results for room reverberations. Only the parameters GOG and SKG are considerable affected by the distortion. The other parameters stay significant at least until a reverberation time of $T_{60} = 130$ ms.

SNR	OQG	GOG	SKG	RCG	IC
∞	**0.000**	**0.014**	**0.000**	**0.006**	**0.000**
50 dB	**0.000**	**0.022**	**0.000**	**0.012**	**0.000**
25 dB	**0.000**	**0.048**	**0.000**	0.708	**0.000**
0 dB	**0.000**	**0.000**	**0.000**	1.000	1.000
-10 dB	**0.000**	**0.000**	**0.000**	1.000	1.000

Table 2. T-test for gender under white noise

SNR	OQG	GOG	SKG	RCG	IC
∞	**0.000**	**0.014**	**0.000**	**0.006**	**0.000**
50 dB	**0.000**	**0.024**	**0.000**	**0.012**	**0.000**
25 dB	**0.000**	**0.032**	**0.000**	0.151	**0.000**
0 dB	0.151	**0.000**	**0.000**	1.000	1.000
-10 dB	0.675	0.060	**0.000**	1.000	0.901

Table 3. T-test for gender under factory noise

T_{60}	OQG	GOG	SKG	RCG	IC
0	**0.000**	**0.014**	**0.000**	**0.006**	**0.000**
60	**0.000**	1.000	0.261	**0.000**	**0.031**
130	**0.003**	1.000	0.827	**0.000**	**0.016**
160	0.083	1.000	0.673	**0.000**	**0.000**

Table 4. T-test for gender under room reverberation

Table 5 presents the number of distinguishable pairs from six voice qualities under white noise disturbance. The total number of pairs is $\binom{6}{2} = 15$. We see that even at an SNR of 0dB, there is still a reasonable number of distinguishable pairs. Figure 3 shows the mean values and the standart deviations of the parameter GOG for different voice qualities. We see that 13 of all 15 pairs except for rough/creaky and normal/falsetto differ in GOG.

Table 6 does the same analysis for seven different emotional states under pink noise. For the noiseless case, high numbers of distinguishable emotional states are noticeable. Like for voice qualities, there is still an appropriate number of distinguishable pairs for emotions, even at an SNR of 0dB. In Figure 4 we see that 20 of 21 pairs of emotions except for fearful/neutral differ in the feature SKG.

5. CONCLUSION

This paper introduced voice quality parameters and demonstrates their potential use to distinguish between speaking

SNR	OQG	GOG	SKG	RCG	IC	total
∞	13	13	9	12	8	55
40 dB	14	13	10	12	10	59
20 dB	13	13	9	12	8	55
0 dB	12	12	10	11	7	52
-20 dB	1	0	0	0	0	1

Table 5. Multiple t-tests for voice qualities under white noise

SNR	OQG	GOG	SKG	RCG	IC	total
∞	20	19	20	19	17	95
40 dB	19	18	21	20	18	96
20 dB	20	18	20	20	17	95
0 dB	16	16	16	15	6	69
-20 dB	0	0	0	0	0	0

Table 6. Multiple t-tests for emotions under pink noise

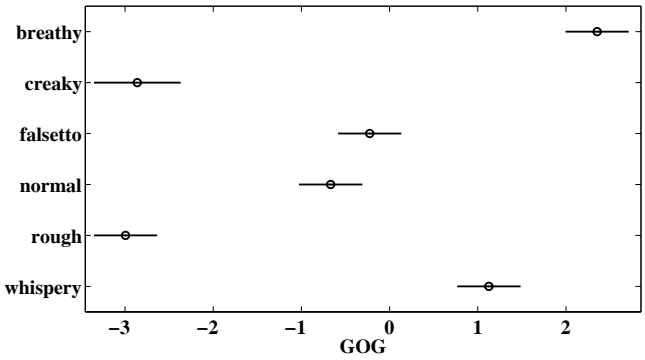

Fig. 3. GOG for six voice qualities (SNR = 20 dB)

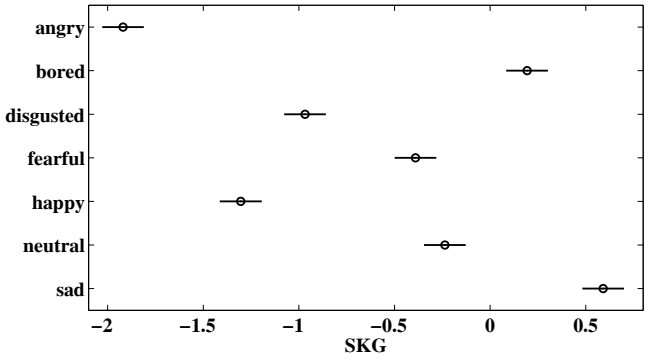

Fig. 4. SKG for seven emotional states (SNR = 20 dB)

groups like gender, voice qualities, and emotional states. It also showed that some of the voice quality parameters are quite robust adverse acoustic background noise and room reverberations. They remain significant in feature distinction under disturbances in a wide range of SNR. This is a first step forward to the long-term objective of building simple but robust classificators for different speaking groups. First classification tests have been done for gender, emotion, and voice quality but could not be presented here due to the limited space of the paper. Some of the promising results are reported in [16].

6. REFERENCES

[1] W. Wokurek and M. Pützer, "Automated corpus based spectral measurement of voice quality parameters," *Proceedings of the International Congress of Phonetic Sciences, Barcelona*, pp. 2173–2176, 2003.

[2] M. Masarek and M. Pützer, "Differenzierung gesunder Stimmqualitäten und Stimmqualitäten bei Rekurrensparese mit Hilfe elektroglottographischer Messung und RBH-System," *Sprache Stimme Gehör*, 2000.

[3] K. Classen, G. Dogil, M. Jessen, K. Masarek, and W. Wokurek, "Stimmqualität und Wortbetonung im Deutschen," *Linguistische Berichte*, vol. 174, pp. 202–245, 1998.

[4] H. Strik, B. Cranen, and L. Boves, "Fitting a LF-model to inverse filter signals," *Eurospeech*, vol. 1, pp. 103–106, 1993.

[5] A. Paeschke, M. Kienast, and W. Sendlmeier, "F0-contours in emotional speech," *Proceedings of the International Congress of Phonetic Sciences*, pp. 929–932, 1999.

[6] A. Paeschke and W. Sendlmeier, "Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements," *Proceedings of the ISCA-Workshop "On Speech and Emotion", Belfast, Nordirland*, pp. 75–80, 2000.

[7] K. Stevens and H. Hanson, "Classification of glottal vibration from acoustic measurements," *Vocal Fold Physiology*, pp. 147–170, 1994.

[8] D. Talkin, W. Kleijn, and K. Paliwal, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis, Elsevier*, pp. 495–518, 1995.

[9] David Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *Technical Report, Bell Labs.*, 1987.

[10] G. Fant, *Acoustic theory of speech production*, The Hague: Mouton, 1970.

[11] M. Pützer and J. Koreman, "A German database of patterns of pathological vocal fold vibration," *PHONUS 3, Universität des Saarlandes*, pp. 143–153, 1997.

[12] J. Laver and H. Eckert, *Menschen und ihre Stimmen - Aspekte der vokalen Kommunikation*, Beltz, 1994.

[13] John Laver, *The phonetic description of voice quality*, Cambridge University Press, 1980.

[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," *to appear in Proceedings of Interspeech*, 2005.

[15] J. Allen and D. Berkley, "Image method for efficiently simulation small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[16] M. Lugger and B. Yang, "Classification of different speaking groups by means of voice quality parameters," *to appear in: ITG Sprachkommunikation 2006*.