Voice Assessment of Speakers with Laryngeal Cancer by Glottal Excitation Modeling Based on a 2-Mass Model

Tobias Bocklet, Elmar Nöth, and Georg Stemmer

Lehrstuhl für Informatik 5 (Mustererkennung)
Friedrich-Alexander-Universität Erlangen-Nürnberg
Martensstr. 3, 91058 Erlangen, GERMANY
tobias.bocklet@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de

Abstract. The paper investigates the automatic evaluation of voice-related criteria of speakers with laryngeal cancer using a parametric two-mass model of the glottis. In contrast to previous approaches based on automatic speech recognition, the proposed method allows for a distinct evaluation of voice parameters alone since the underlying feature extraction technologies are based on a modeling of the whole vocal tract. This work focuses on the separation of vocal folds and vocal tract by LPC, where the vocal folds are represented by a parametric two-mass model which characterizes the excitation signal. The model parameters are optimized by a data-driven optimization procedure in order to fit the synthetic excitation signal to the LPC residue and the estimated pitch. We found first evidence that the computed parameters are meaningful in form of Pearson correlations between excitation signal parameters and different perceptual voice evaluation criteria in the range of $r \approx |0.7|$.

Keywords: Glottal excitation, voice modeling, perceptual evaluation.

1 Introduction

Voices of speakers with partial laryngectomy are often more hoarse, more harsh and more aspirated than normal speakers [11,7]. This can be explained by the anatomic alterations due to the cancer and/or the following treatment which may lead to a restricted movement of the vocal folds or to an inaccurate closure of the vocal folds [3,5,9]. In clinical routine, the quality of a person's voice is perceptually evaluated with respect to different rating criteria. Modeling the vocal folds by an adequate physical model may allow an automatic evaluation of distinct voice parameters.

This work focuses on the automatic evaluation of voice-related criteria on the basis of connected speech (read texts). In order to allow a (distinct) analysis of voices, an approach that is based on the source-filter model of the speech generation process is employed for voice evaluations. Voiced speech sounds are generated by the excitation signal, i.e., the source signal of the glottis. This signal is filtered by the vocal tract, where different frequencies are amplified or softened. In order to allow a meaningful evaluation of a person's voice, the influence of the vocal tract has to be omitted. This is achieved by assuming a linear filter between glottis and vocal tract. Linear prediction is

I. Habernal and V. Matoušek (Eds.): TSD 2011, LNAI 6836, pp. 348–355, 2011.

applied to obtain the vocal tract configuration for each time frame. As an approximation of the excitation signal, the residue of the Linear Predictive Coding (LPC), an inverse filtering of the speech signal with the LPC filter is calculated in an data-driven optimization procedure. The model parameters are now optimized to match the synthetic excitation signal as close as possible to the LPC residue and the estimated pitch. The final parameters are then analyzed with respect to different voice evaluation parameters.

Automatic assessment of voice and speech criteria has already been investigated in previous works using different automatic speech processing techniques. [8], for instance, shows high correlations of different articulation and voice criteria between perceptual ratings and an automatic speech recognition (ASR) when a standard text is spoken. The usage of ASR systems for intelligibility assessment can be easily motivated: If the intelligibility of a speaker is low, the word recognition rate is low. A possible disadvantage of approaches based on ASR is that they ”=in contrast to the method presented in this paper”= account for the complete speech signal in form of spectral features. These features contain information of both the excitation signal of the vocal folds and the formant structure of the vocal tract/acoustic tube. This is not always intended, especially when it comes to the evaluation of distinct voice-related aspects.

The outline of this work is as follows: We first describe the used data in Chapter 2. The basics of the glottal excitation system are given in Chapter 3, the results are discussed in Chapter 4. The work is concluded by a summary and a short outlook in Chapter 5.

2 Dataset

Audio data were recorded from 85 patients (75 men, 10 women) suffering from cancer in different regions of the larynx. 65 of them had already undergone surgery with partial laryngectomy. They have been recorded 2.4 months after surgery on average. 20 speakers were still awaiting surgery. The average age of all speakers was 60.7 ±9.7 years. The youngest and the oldest person were 34 and 83 years old, respectively. Fig. 1 shows a patient before and after treatment of a T1 tumor of the right vocal fold. Before treatment the vocal fold oscillation of the right vocal fold is strongly limited. After surgery and radiotherapy the tumor is eliminated and the vocal fold oscillation has improved.

Each person read the text “Der Nordwind und die Sonne”, a phonetically balanced text with 108 words (71 disjunctive) which is used in German speaking countries in speech therapy. The English version is known as “The North Wind and the Sun” [4]. The speech data were sampled with 16 kHz and an amplitude resolution of 16 bit.

In order to obtain references for the automatic evaluation, five experienced phoniatricians and speech scientists evaluated each speaker regarding different voice (and speech) criteria. Voice quality, penetration, tone and intelligibility were rated regarding a 5-point scale with the labels very high, high, moderate, low, and none. Each raters decision for each patient was converted to an integer number between 1 and 5. Additionally roughness (R), breathiness (B) and hoarseness (H) were rated regarding the RBH-scale [13]. Evaluations are quantized from 0 (not present) to 3 (intense).

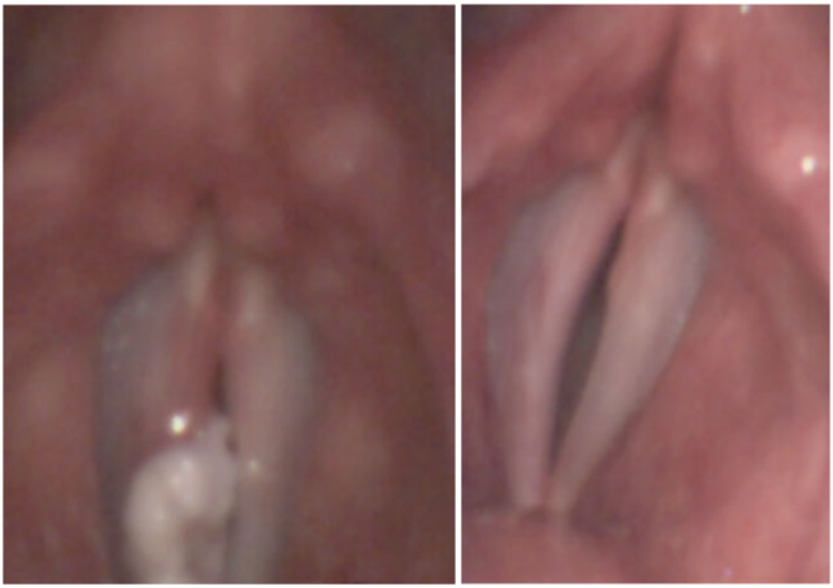

Fig. 1. Example of a person with a T1 tumor on the right vocal fold (left picture). After surgery and radiotherapy the tumor is eliminated and the vocal fold oscillation has improved (right picture). Oscillation capability has not recovered completely.

3 Glottal Excitation

3.1 Two Mass Model

The approach estimates the parameters of a physical glottis model from data of speakers with laryngeal cancer. The goal is to find pathology-related changes in the model parameters that reflect the voice quality and other voice related evaluation criteria. Therefore, the used glottis model should ideally have physically meaningful parameters, in contrast to just describing the shape of the excitation signal. The model should be flexible enough to adequately represent pathology-related changes of the voice quality.

Considering these requirements we employed the two-mass vocal fold model introduced by Stevens [12] and illustrated in Fig. 2. The model consists of two pairs of masses, larger ones (M_1) representing the inferior part of the vocal folds, and small ones (M_2) representing the superior part of the vocal folds. The model is symmetrical, there is no differentiation between the masses of the left and right side. The mechanism depends on the fact, that the inferior and superior part of the vocal folds do not move together as a rigid body. There is a certain degree of freedom to move relatively to each other [2]. This freedom is modeled by a coupling compliance by springs. Each mass moves on a spring that is connected with the latter wall. The masses are connected among themselves by an additional spring. The compliances of the springs are described by the parameters C_1, C_2 and C_c (for the spring that connects M_1 with M_2). Note that parameters for the masses and compliances are given as *mass per unit length* and

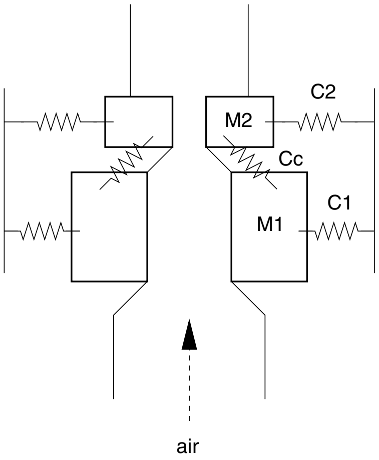

Fig. 2. Two-mass vocal fold model by Stevens [12]

compliance per unit length, i.e., they may change when the vocal folds are stretched. Air flows from bottom to top through the glottis when both M_1 and M_2 have a positive displacement, as shown in Fig. 2.

The excitation function of the two-mass vocal fold model by Stevens is obtained in three steps. First, the displacements $x_1(t)$ and $x_2(t)$ of the inferior and superior part of the vocal folds over time t are computed. The width of the glottal opening $d(t)$ is defined to be $\min(x_1(t), x_2(t))$. Second, from the width of the opening, the airflow $U_g(t)$ through the glottis is determined. In the third step, taking the derivative of $U_g(t)$ results in the excitation function.

The whole process of the excitation function computation is described in Chapter 2 of [12]. However, some details cannot be found in the book. In [1] a detailed derivation of all model formulas is given. The initial and fixed values for all parameters are taken from [12] and summarized in [1].

3.2 Model Optimization

Our hypothesis is that glottis model parameters contain information about the degree of pathology of speakers with laryngeal cancer. To test this hypothesis, we find the optimal model parameters that fit the speech data and observe how they change with varying pathology.

Figure 3 depicts a block diagram of the optimization loop. A set of initial parameters (M_1, M_2, C_1, C_c, x_0, d_1, ϕ, l) is the input of the glottis excitation model. The model generates an excitation signal for a 10 ms speech frame. At the same time, the LPC residue of the original speech signal is calculated and the log spectrum transform is applied to both of these excitation signals. The similarity of the generated excitation signal is compared to the original signal using two Euclidean distances. The distance between the log spectrum of the two signals is compared in a first step. In a second step, the distance between the generated and the original pitch for the frame are compared.

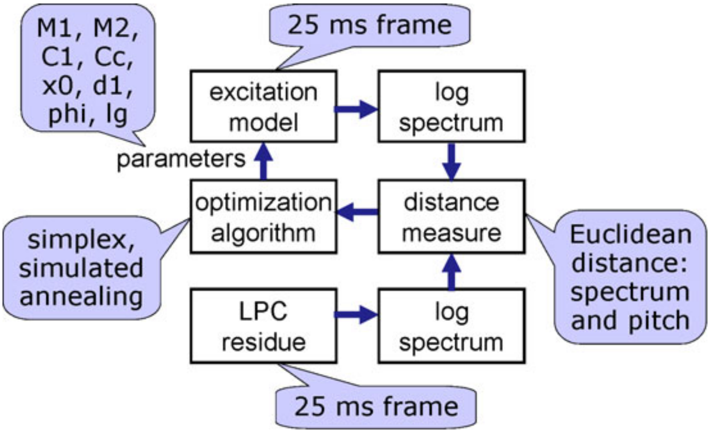

Fig. 3. Optimization of the parameters of the glottal excitation model

The combined distance measure is passed to the optimization algorithm, which modifies the parameter set, passing the new parameter set to the excitation model. Thus, an optimization loop is formed, modifying the parameters, generating a new candidate excitation signal, and testing it against the original signal. The simplex algorithm [10] and simulated annealing [6] are used for optimization.

The optimization is formulated as:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}[D(s_m(\theta), s_{org})] \tag{1}$$
$$\theta = \{M_1, M_2, C_1, C_c, x_0, d_1, \phi, l\}$$

where $D(s_m(\theta), s_{\mathrm{org}})$ is the combined distance between the model excitation signal s_m and the original excitation signal s_{org}.

The combined distance measure combines distances between both the respective log spectra and the respective pitches p_m, p_{org}, and is defined as:

$$D(s_m(\theta), s_{\mathrm{org}}) = D(\mathrm{logspec}(s_m(\theta)), \mathrm{logspec}(s_{\mathrm{org}})) + \lambda \cdot D(p_m, p_{\mathrm{org}}) \tag{2}$$

where $D(\cdot,\cdot)$ is the Euclidean distance between two vectors and the constant λ scales the influence of the pitch distance. Note that the optimization is only performed on voiced speech segments.

Table 1 contains a description of all parameters of the glottal excitation model. The exact derivation of the formulas and parameters is omitted here. For description see [12] and [1].

The excitation parameters (Table 1) are calculated every 10 ms. We assumed that the standard deviations of the parameters, calculated per speaker over the whole text, are meaningful features for voice evaluations. In preliminary experiments we confirmed this. Pearson correlation coefficients are used as agreement measure.

Table 1. Description of the parameters of the glottal excitation model

param	description
M_1	mass of inferior part of the vocal fold
M_2	mass if superior part of the vocal fold
C_1	compliance of spring between M1 and lateral wall
C_c	compliance of spring between M_1 and M_2
d_1	average vertical length of the lower portion of the vocal fold
x_0	resting position of M_1 in the absence of any force
ϕ	skewness factor; representation of the constriction of the vocal tract
l	length of glottis (assuming rectangular shape)
D	optimization distance measure (see Eq. 2)

4 Results and Discussion

In Table 2 the Pearson correlation coefficients among the different evaluation criteria is given. Pearson coefficients of $r > 0.9$ are measured between voice quality and tone, between tone and intelligibility, and between intelligibility and voice quality. Voice quality is highly connected to penetration, breathiness and hoarseness. The same statement holds also for tone and intelligibility. Among the RBH-scale breathiness and hoarseness and roughness and hoarseness correlate with $r > 0.8$. Roughness achieves only moderate ($r < 0.65$) correlations to voice quality, penetration, tone and intelligibility.

Table 2. Pearson correlation among the different perceptual evaluation criteria voice quality (quality), voice penetration (penetr), tone, intelligibility (intell), roughness (R), breathiness (B), hoarseness (H)

	penetr	tone	intell	R	B	H
quality	0.87	0.93	0.90	0.63	0.82	0.84
penetr		0.84	0.86	0.45	0.73	0.70
tone			0.93	0.64	0.84	0.85
intell				0.59	0.83	0.80
R					0.56	0.84
B						0.81

The standard deviation of the excitation parameters are compared with the mean values of the perceptual evaluation criteria. The results are summarized in Table 3. We achieved moderate to good Pearson coefficients. Note that all of the correlation coefficients in Table 3 are negative. That means for example speaker with a high voice quality have a high standard deviation in masses M_1 and M_2. Note that the 9 excitation parameters are calculated every 10 ms. The variation between these 10 ms segments, i.e., phonemes, is high for speakers with a high voice quality. The changes of the parameters are lower between the 10 ms segments, when the speakers have a lower voice quality.

Voice quality correlates with $r = -0.69$ and $r = -0.67$ to the two masses M_1 and M_2. The Pearson coefficient between voice quality and the compliance of the spring between M_1 and M_2 (C_c) is $r = -0.71$. These three parameters achieve correlation coefficient in the same order for the criteria *tone* and *intelligibility*. This result is not really surprising, since these criteria correlate highly among each other (see Table 3).

The excitation parameter ϕ that represents the constriction of the vocal tract achieves a Pearson coefficient of $r = -0.69$ with the criterion *voice penetration*. People with a good *voice penetration*, have a high standard deviation of ϕ, the constriction of the vocal tract changes a lot. This can be explained by a strong variation of the air flow. Note, that ϕ reaches such a high correlation only for the criterion penetration, for all other criteria it seems not to be an adequate feature.

The excitation parameters M_1, M_2 and C_c achieve the highest correlation coefficients for the criteria of the RBH-scale. The differences in correlation coefficients of these three excitation parameters are not significant for these three criteria, nevertheless the mass M_1 achieves slightly better results. Breathiness achieved $r = -0.70$ and hoarseness achieved $r = -0.65$. Roughness showed only moderate correlations of $r = -0.41$. The length l of the glottis shows moderate correlations $r \approx 0.6$ for most criteria. The parameters D, d_1, x_0 achieved only moderate correlations.

Table 3. Pearson correlation results between the perceptual evaluation criteria and the standard deviation of the parameters of the excitation system. The highest Pearson coefficient for each evaluation criterion is marked bold.

param	quality	penetr	tone	intell	R	B	H
M_1	-0.69	-0.62	-0.71	-0.63	**-0.41**	**-0.70**	**-0.65**
M_2	-0.67	-0.60	-0.69	-0.61	-0.39	**-0.70**	-0.63
C_1	-0.50	-0.39	-0.54	-0.42	-0.30	-0.59	-0.48
C_c	**-0.71**	-0.66	**-0.72**	**-0.65**	-0.40	-0.68	**-0.65**
ϕ	-0.54	**-0.69**	-0.51	-0.51	-0.23	-0.44	-0.42
l	-0.61	-0.54	-0.61	-0.57	-0.34	-0.55	-0.59
D	-0.53	-0.44	-0.54	-0.47	-0.27	-0.54	-0.49
d_1	-0.45	-0.38	-0.48	-0.44	-0.22	-0.45	-0.41
x_0	-0.24	-0.23	-0.27	-0.27	-0.07	-0.20	-0.22

5 Summary

In this work we applied a newly-developed glottal excitation system to the task of voice evaluations of speakers with laryngeal cancer. The system adapts different glottal parameters in a data-driven optimization loop to speech frames of 10 ms. We showed correlations between different parameters of the excitation system and speech evaluation criteria. The two masses and the compliance of the spring between these two masses showed good correlations to the parameters voice quality, penetration, tone, intelligibility breathiness and hoarseness. The parameter ϕ that represents the constriction of the vocal tract, showed the best correlation to the criterion penetration. In future work we plan to adapt more complex vocal fold models in order to achieve higher agreement between the model parameters and perceptual evaluations. Examples are the use of a non-symmetrical vocal fold model or pitch synchronous modeling.

Acknowledgments. The algorithm for the parameter estimation of the glottis model has been developed and implemented by Georg Stemmer and Andrew Cassidy at the Johns-Hopkins-Summerworkshop 2008. This study was partially funded by the German Research Foundation (DFG) under grant EY 15/18-2. We thank the speech therapists of our department for the perceptual evaluations.

References

1. Beyerlein, P., Cassidy, A., Kholhatkar, V., Lasarcyk, E., Nöth, E., Potard, B., Shum, S., Song, Y.C., Spiegl, W., Stemmer, G., Xu, P.: Vocal aging explained by vocal tract modelling: 2008 JHU summer workshop final report. Tech. rep (2008)
2. Fant, G.: Acoustic Theory of Speech Production. Mouton, Netherlands (1960)
3. Fung, K., Lyden, T., Lee, J., Urba, S., Worden, F., Eisbruch, A., Tsien, C., Bradford, C., Chepeha, D., Hogikyan, N., Prince, M., Teknos, T., Wolf, G.: Voice and swallowing outcomes of an organ-preservation trial for advanced laryngeal cancer. Int. J. Radiat. Oncol. Biol. Phys. 63(5), 1395–1399 (2005)
4. Handbook of the International Phonetic Association. Cambridge University Press, Cambridge (1999)
5. Kim, C., Lim, Y., Kim, K., Kim, Y., Choi, H., Kim, K., Choi, E.: Vocal analysis after vertical partial laryngectomy. Yonsei. Med. J. 44(6), 1034–1039 (2003)
6. Kirkpatrick, S., Gelatt, C., Vecchi, M.: Optimization by Simulated Annealing. Science 220(4598), 671–680 (1983)
7. Kosztya-Hojna, B., Rogowski, M., Pepiski, W., Rutkowski, R., Lazarczyk, B.: Voice analysis after the partial laryngectomy in patients with the larynx carcinoma. Folia histochemica et cytobiologica Polish Academy of Sciences Polish Histochemical and Cytochemical Society 39(Suppl 2), 136–138 (2001)
8. Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., Nöth, E.: PEAKS - A system for the automatic evaluation of voice and speech disorders. Speech Communication 51(5), 425–437 (2009)
9. Makeieff, M., Barbotte, E., Giovanni, A., Guerrier, B.: Acoustic and aerodynamic measurement of speech production after supracricoid partial laryngectomy. Laryngoscope 115(3), 546–551 (2005)
10. Nelder, J.A., Mead, R.: A Simplex Method for Function Minimization. The Computer Journal 7(4), 308–313 (1965)
11. Olthoff, A., Mrugalla, S., Laskawi, R., Fröhlich, M., Stürmer, I., Kruse, E., Ambrosch, P., Steiner, W.: Assessment of irregular voices after total and laser surgical partial laryngectomy. Arch. Otolaryngol Head Neck Surg. 129(9), 994–999 (2003)
12. Stevens, K.N.: Acoustic Phonetics. The MIT Press, Cambridge (1998)
13. Wendler, J., Rauhut, A., Krüger: Classification of voice qualities. Journal of Phonetics 14, 483–488 (1986)