CLASSIFICATION OF DIFFERENT SPEAKING GROUPS BY MEANS OF VOICE QUALITY PARAMETERS

M. Lugger and B. Yang

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

ABSTRACT

This paper presents a new method to classify different speaking groups by using the so called voice quality parameters. By voice quality we mean the characteristics of the glottal excitation of the speech signal. We estimate five parameters describing the voice quality as spectral gradients of the vocal tract compensated speech. They correlate with the glottal speech source features like open quotient, skewness of the glottal pulse, and incompleteness of closure. These parameters are then used to classify gender, different phonation types, and different emotions with promising results.

1. INTRODUCTION

Detection of paralinguistic properties of speech is in contrast to other applications in speech processing like automatic speech recognition a less explored field. Under paralinguistic properties we understand all the information beyond to the pure linguistic content of a spoken utterance. They describe the emotional state of a speaker, the peculiarities or manner of his individual voice, his gender, or his social background. So the listener obtains information about the physical, psychological, social, and emotional characteristics of the speaker.

According to the source-filter-model, the linguistic content of speech is mainly determined by the vocal tract while the glottal excitation contributes significantly to the paralinguistic properties. We call the characteristics of the glottal excitation as voice quality. It is our goal to extract parameters describing the voice quality from the speech signal which allow a mapping of spoken utterances to different speaking groups.

This paper shows three classification applications of the voice quality parameters: the gender, the phonation types by J. Laver [1], and emotional states. The paper is structured as follows. Section 2 explains how the voice quality parameters are estimated. In section 3, the speech data and the classification method is described. The results of the classifications are presented in section 4.

2. VOICE QUALITY PARAMETERS

Voice quality is mainly affected by the excitation of the human voice that is called phonation. That means, the shape of the glottal pulse and its rate and time variations are responsible for the kind of voice quality that is realized. In contrast, all activities belonging to the articulation process affect the sounds that all together build the content of the speech. In the literature, some non-acoustic methods measure the change in electrical impedance across the throat during speaking. An electroglottograph, for example, measures the change in electrical impedance across the throat during speaking.

We study a method to measure the voice quality 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 [2] 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.

We use a modified algorithm which calculates spectral gradients instead of amplitude ratios. In addition, a vocal tract compensation is performed prior to estimating the gradients [3]. The whole algorithm 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.

2.1. 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 [4] 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 [5]. All frequency values are converted to the Bark scale. Table 1 shows the estimated speech features required for the voice quality estimation.

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	amplitudes at F_p and $2F_p$ [dB]
F_{1p}, F_{2p}, F_{3p}	frequencies of peaks near formants
A_{1p}, A_{2p}, A_{3p}	amplitude at F_{1p}, F_{2p}, F_{3p} [dB]

Table 1. Speech features used for voice quality estimation

2.2. 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 [6]

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

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

$$\begin{aligned} \tilde{H}_k &= H_k - \sum_{i=1}^{4} V_{dB}(kF_p; F_i, B_i) \qquad (k = 1, 2) \\ \tilde{A}_{kp} &= A_{kp} - \sum_{\substack{i=1 \\ i \neq k}}^{4} V_{dB}(F_{kp}; F_i, B_i) \qquad (k = 1, 2, 3) \end{aligned}$$

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.

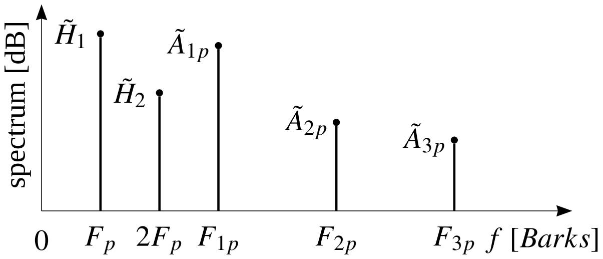

Fig. 1. Vocal tract compensated peaks of the FFT spectrum for the voice quality parameter estimation

2.3. Estimation of the voice quality parameters

The last step estimates the following five voice quality parameters from the vocal tract compensated speech features: ”Open Quotient Gradient”, ”Glottal Opening Gradient”, ”SKewness Gradient”, ”Rate of Closure Gradient”, and ”Incompleteness of Closure”. 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.

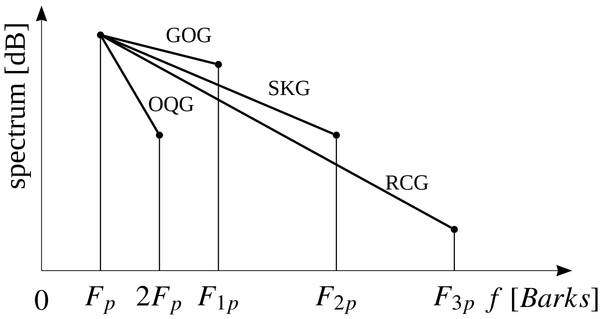

Fig. 2. Voice quality parameters as spectral gradients

3. EXPERIMENTS

In this paper, we apply the voice quality parameters to classify different speaking groups in three studies: classification of gender, voice qualities, and emotions. The speech data for all studies are recorded in an anechoic room at the sampling frequency of 16 kHz. All signals were segmented to utterances of about 3 second length. The speech was classified on the basis of every spoken utterance.

3.1. Classification

There are many different approaches for gender and emotion detection in the literature. A method for gender detection is described in [7]. A good overview of previous works about emotion detection is given in [8]. Most of the proposed methods use prosodic properties of the speech like the intonation, intensity, and duration as features for the classification. They do not consider voice quality parameters as features because they found them difficult to model and to estimate. Similarly, there exist many different methods for pattern recognition. For emotion detection, the Bayes classifier, Gaussian mixture modes, hidden Markov models, and artificial neural networks have been used.

In this paper, we use voice quality parameters only for all classifications. For the pattern recognition part, we use a linear discriminant analysis because it is simple and there is no need for a long training phase.

3.2. Linear discriminant analysis

The discriminant analysis is a method for multivariate data analysis [9]. In this work, a linear discriminant analysis is used for classification. It involves several linear discriminant functions.

In the training phase, the parameters of the linear discriminant functions are determined from a random subset of the speech data with a priori known classes. This is called supervised training. These discriminant functions are then used in a second test phase to classify the remaining(or complete) speech data into different speaking groups. The results of the classifier are compared with the a priori known classes in order to evaluate the quality of the classifier.

4. RESULTS

This section presents the classification results. Except for the gender, all classifications are speaker dependent, yet. The classification was done for every spoken utterance by a majority decision. That means, every voiced segment of an utterance is mapped to one class by the classifier. The whole utterance is classified to that class with the maximum of single segment decisions. The confusion matrices for the three studies are depicted in the Table 2-4. The first column contains the true speaking group and the first line shows the detected speaking group. The entries in boldface on the main diagonal show the rates of the correct decisions. The other non-diagonal entries are the rates of wrong decisions.

4.1. Classification of gender

The speech samples for the first study were taken from [10]. They consist of 10 utterances of male and 10 utterances of female speech, each of about 30 second duration. For the training phase a random set of 5 male and 5 female speakers was considered. Based on these training data a classification of all 20 speakers was performed.

Table 2 shows the gender classification under white background noise with different values for the global signal to noise ratio (SNR). We see that the gender classification shows quite good results. Only 7.4% of the male respectively 4.8% of the female utterances are wrong classified in the noiseless case. For a decreasing SNR, there is a moderate and smooth performance degradation where the classifier tends to favour the female decisions. The reason is that the female voice has a stronger breathy portion as the male voice [11] and hence noisy speech tends to be classified to femal than to male.

gender	SNR	male	female
male	∞	**92.6%**	7.4%
female		4.8%	**95.2%**
male	30 dB	**82.7%**	17.3%
female		9.3%	**90.7%**
male	15 dB	**81.5%**	18.5%
female		11.3%	**88.7%**
male	0 dB	**58.5%**	41.5%
female		5.6%	**94.4%**

Table 2. Gender classification under white noise

4.2. Classification of phonation types

The noiseless speech data for the second study were taken from the book [12] by J. Laver. It contains utterances spoken in six different phonation types [1] by the same speaker: ”modal voice”, ”falsetto voice”, ”whispery voice”, ”breathy voice”, ”creaky voice”, and ”rough voice”. For the classification, only the phonation types ”rough voice”, ”creaky voice”, and ”modal voice” were considered.

Table 3 shows the results of the phonation type detection. The classification for modal voice shows a very good result with a detection rate of 95.6%. Rough voice is correctly detected in nearly 3 out of 4 utterances. Creaky voice is correctly detected in 67.5%. It is mainly confused with the modal voice.

voice quality	modal	rough	creaky
modal voice	**95.6%**	0.0%	4.4%
rough voice	12.7%	**73.3%**	14.0%
creaky voice	28.8%	3.7%	**67.5%**

Table 3. Classification of phonation types

4.3. Classification of emotions

The noiseless speech data for the third study were taken from the ”Berlin database of emotional speech” [13]. It contains about 500 sentences spoken by actors in a neutral, happy, angry, sad, fearful, bored, and disgusted way. We used the emotions angry, sad, happy, and neutral only. The classification was done speaker-by-speaker separately. That means, speech samples of the same speaker were used for both the training and classification. The values in the confusion matrix are mean values over 10 speakers.

Table 4 shows the results of emotion detection. We see that the emotions angry, sad, and neutral were classified with detection rates over 80%. Only happy voice shows a detection rate of 57.7% because it is often confused with angry voice. This is a well known fact from the literature [14]. Happy and angry voices show similar values in the emotion dimension activity and valency. They only differ in the dimension potency.

Emotion	angry	sad	happy	neutral
angry	**93.7%**	0.0%	5.6%	0.8%
sad	0.0%	**84.7%**	1.7%	13.6%
happy	33.8%	0,0%	**57.7%**	8.4%
neutral	0.0%	2.5%	6.3%	**91.1%**

Table 4. Classification of emotions

5. CONCLUSION

We introduced the voice quality to describe the glottal excitation and presented an algorithm to estimate the voice quality parameters from the speech signal. The classification of gender, phonation type, and emotion by using voice quality parameters shows promising results. The next step is to achieve a speaker independent classification for phonation type and emotion. One idea could be the use of additional prosodic features like pitch, intensity, and duration. A second approach could be the improvement of the classifier.

6. REFERENCES

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

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

[3] M. Pützer and W. Wokurek, “Multiparametrische Stimmprofildifferenzierung zu männlichen und weiblichen Normalstimmen auf der Grundlage akustischer Signale,” *Laryngo-Rhino-Otologie, Thieme*, 2006.

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

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

[6] G. Fant, *Acoustic theory of speech production*, The Hague: Mouton, 1960.

[7] H. Harb and L. Chen, “Gender identification using a general audio classifier,” *Proc. IEEE International Conference on Multimedia and Expo,*, pp. 733–736, 2003.

[8] R. Cowie and E. Douglas-Cowie, “Emotion recognition in human-computer ineraction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[9] R. O. Duda, P. Hart, and D.G. Stork, *Pattern Classification*, Wiley, 2001.

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

[11] H. Hanson and E. Chuang, “Glottal characteristics of male speakers: Acoustic correlates and comparison with female data,” *Journal of acoustical society of America*, vol. 106, pp. 1064–1077, 1999.

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

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” *Proceedings of Interspeech*, 2005.

[14] Astrid Paeschke, *Prosodische Analyse emotionaler Sprechweise*, Ph.D. thesis, TU-Berlin, 2003.

[15] M. Lugger, B. Yang, and W. Wokurek, “Robust estimation of voice quality parameters under real world disturbances,” *In: Proc. IEEE ICASSP*, 2006.