PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Characterization of turbulence noise in breathy human phonation

Philipp AICHINGER

Medical University of Vienna, Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics, Austria

ABSTRACT

Breathiness is a voice quality type that may be a sign of a voice disorder. It involves the auditory perception of additive noise caused by turbulent trans-glottal airflow. Ten normal and ten breathy phonations recorded during high-speed videolaryngoscopy are analyzed. A method for extracting additive noise from audio recordings in the presence of modulation noise is presented. It generates quasi-unit pulse trains at a rate equal to the vocal frequency. The cycle shape is obtained by cross-correlating the pulse train with the audio signal. The cycle shape is then Fourier transformed and input to a Fourier synthesizer. Other inputs are the estimates of the instantaneous phase and the amplitude modulation function. The timing and height of the pulses are modulated so as to minimize the modeling error, which is the difference between the recorded audio signal and the output of the synthesizer. The error is assumed to approximate the additive noise in the voice. Differences of energy levels and spectral slope with respect to voice quality (breathy/normal) are reported. No differences with regard to cycle phase (open/quasi-closed) are reported.

Keywords: Dysphonia, breathiness, laryngeal high-speed videos, voice quality typing

1. INTRODUCTION

Voice quality typing is relevant to the clinical care of voice disorders, because it contributes to the indication, selection, evaluation, and optimization of clinical treatment. A particular voice quality type is breathiness, which is often described as the auditory perception of turbulence noise caused by the trans-glottal air flow. Typically, it is modelled by random noise that is added to the cyclic component of the voice source signal, which is caused by the vibration of the vocal folds. Cycle length and cycle peak modulations, called vocal jitter and shimmer, are frequent in pathological voices. In contrast to additive noise, which is perceived as breathiness, random modulation noise is perceived as roughness (1).

Additive noise observed in pathological voices is often modelled as white Gaussian noise that is low-pass filtered with a spectral slope of –6 dB/octave (2). The filtered noise is multiplied by $h_1 g(n) + h_2$, where $g(n)$ is the magnitude of the trans-glottal airflow rate, and h_1 and h_2 are constants that fix the sizes of the pulsatile and stationary noise components. The stationary component is associated with a residual glottal gap owing to a lack of full contact between the vocal folds during the closed glottis phase. A residual glottal gap is frequently observed in pathological voices. The pulsatile noise component is caused by the cyclic evolution of the glottal area and the trans-glottal airflow.

The purpose of the presentation is to test assumptions regarding additive noise in pathological voices. A signal processing method segregating additive and modulation noise in voiced speech signals is presented first. The additive noise is described. Energy levels, power spectral densities and probability density functions of the estimated additive noise are reported with respect to voice quality (breathy/normal) and glottal phase (open/quasi-closed).

2. MATERIAL AND METHODS

2.1 Test signals

Ten breathy and ten normal sustained phonation signals are selected for testing from a database comprising laryngeal high-speed videos and simultaneous audio recordings of 80 pathological and 40 healthy subjects (3). The subjects are instructed to sustain an /i/ vowel the production of which tilts the epiglottis so as to facilitate sighting of the vocal folds. The phonations sound schwa-like because of the artificially lowered position of the tongue, which co-occurs with rigid video-endoscopy. The test

signals are between 0.256 and 1.872 sec long (mean 0.786 sec).

Figure 1 shows a midsagittal view of the procedure. For video recordings, a Richard Wolf HRES Endocam 5562 is used with a frame rate of 4 kHz. The spatial resolution is 256x256 pixels (interpolated from red channel: 64x128, green channel: 128x128, blue channel: 64x128). For audio recordings, a headworn microphone AKG HC 577 L, a phantom power adapter AKG MPA V L, and a portable audio recorder TASCAM DR-100 are used. The sampling rate is 48 kHz, and the quantization bit resolution is 24 bits. The uncompressed PCM/WAV file format is used. The microphone is used with the original cap (no boost of high frequencies), and the adapter is set to linear response. The low cut of the audio recorder is switched off.

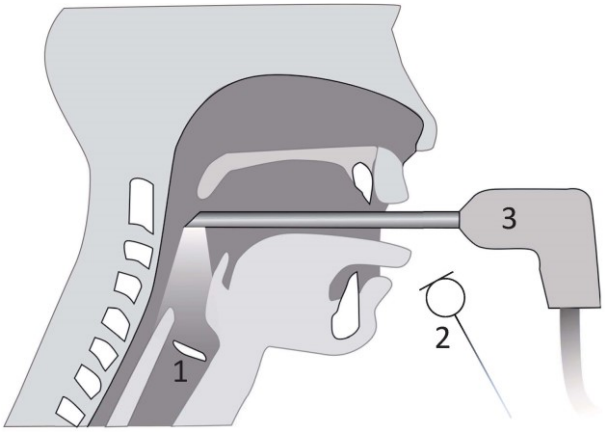

Figure 1 – Midsagittal illustration of the recording procedure. The endoscopic camera is inserted into the mouth of the subject. The vocal folds are illuminated and filmed via the endoscope. A microphone is used to record radiated sound pressure. 1. Vocal folds, 2. Microphone, 3. Endoscopic camera. Figure adapted from (4).

2.2 ESTIMATION OF THE VOCAL FREQUENCY

Figure 2 shows the block diagram of the estimation of the vocal frequency from the audio signals as described in (5) and (6), with a few modifications. First, candidate vocal frequencies f_0^γ are obtained from the speech signal $d(n)$ via spectral peak picking (SPP), and applying the Viterbi algorithm repetitively (six times). γ is the index of the candidate, and n is the discrete time index. Second, a cyclic waveform $d^\gamma(n)$ is obtained for each candidate as follows. A unit-pulse train $u^\gamma(n)$ with frequency f_0^γ is cross-correlated with the audio signal to obtain the cycle shape $r^\gamma(l)$, where l is the discrete lag time index with respect to the center of the cycle. Fourier coefficients a^γ and b^γ are obtained via discrete Fourier transformation (DFT) of the cycle shape $r^\gamma(l)$. The cyclic candidate waveform $d^\gamma(n)$ is obtained via Fourier synthesis (FS), taking f_0^γ, a^γ, and b^γ as inputs. Finally, the vocal frequency estimate f_0 is the candidate which minimizes the difference $e(n) = d(n) - d^\gamma(n)$ in a least squares sense.

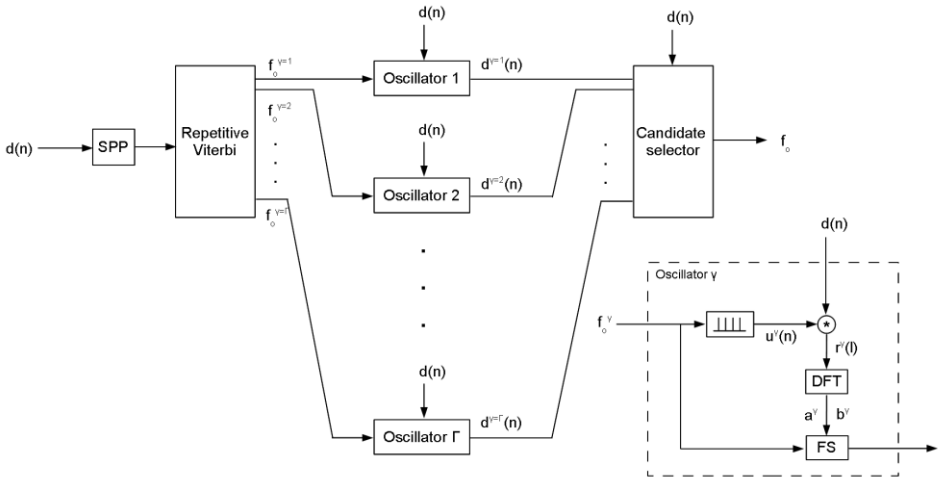

Figure 2 – Block diagram of the vocal frequency estimator, adapted from (6).

2.3 EXTRACTION OF ADDITIVE NOISE IN THE PRESENCE OF MODULATION NOISE

Additive noise is estimated by subtracting a model $\tilde{d}(n)$ of the modulated cyclic component from the speech signal. The modulated cyclic component is obtained as follows. First, a quasi-unit, i.e., an amplitude and frequency modulated pulse train $\tilde{u}(n)$ with average frequency f_0 is generated and cross-correlated with the audio signal $d(n)$ to obtain the cycle shape $\tilde{r}(l)$. Fourier coefficients $\tilde{a}$ and $\tilde{b}$ are obtained by DFT of cycle shape $\tilde{r}(l)$. Second, the instantaneous phase $\widetilde{\Theta}(n)$ is estimated from the pulse train $\tilde{u}(n)$ by assigning $\widetilde{\Theta}(n) = \pi,\ 3\pi,\ 5\pi, \ldots$ to the pulse positions and cubic spline interpolating at the time instants in between. Third, the amplitude modulation function $\widetilde{\mathrm{A}}(n)$ is estimated from pulse train $\tilde{u}(n)$ by assigning $\widetilde{\mathrm{A}}(n) = \tilde{u}(n)$ at the pulse positions and shape-preserving parabolic interpolating at the time instants in between. The modulated cyclic component $\tilde{d}(n)$ is estimated via Fourier synthesis taking as inputs coefficients $\tilde{a}$, $\tilde{b}$, and phase $\widetilde{\Theta}(n)$. The output of the Fourier synthesizer is multiplied by amplitude modulation function $\widetilde{\mathrm{A}}(n)$. Finally, the noise $\tilde{e}(n)$ is minimized in a least squares sense via trial modifications of the pulse positions and heights in train $\tilde{u}(n)$. Deviations of timing from periodicity, $j(\mu)$ and heights from unity, $s(\mu)$ are estimated using the interior-point optimization algorithm (7,8).

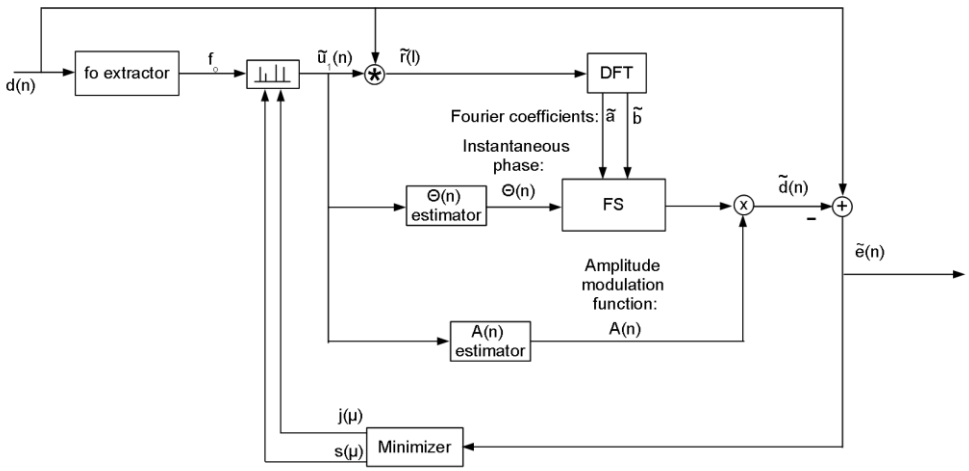

Figure 3 – Block diagram of additive noise extraction, adapted from (6).

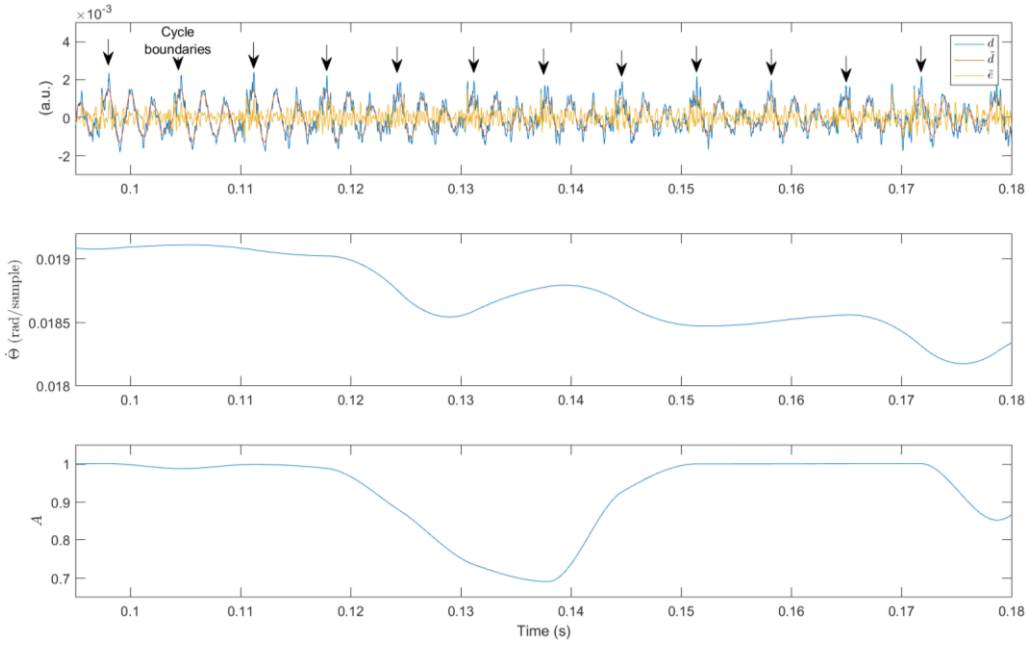

Figure 4 – Fragment of a speech signal. The overlaid plots in the top panel show the audio signal d, the estimate of the cyclic component $\tilde{d}$ and the noise estimate $\tilde{e} = d - \tilde{d}$. The middle and bottom panels show the instantaneous frequency and the amplitude modulation functions of the cyclic component.

2.4 CHARACTERIZATION OF THE ADDITIVE NOISE

Additive noise $e(n)$ is described as follows. First, the noise energy level L_e is obtained in dB relative to the energy level of the cyclic component. Level L_e is reported with respect to voice quality (breathy/normal) and glottal phase (open/closed). The glottal phase is obtained via analysis of the glottal area observed in the video images, taking into account a delay owing to the time of sound propagation through the vocal tract. For comparison purposes, the Harmonics-to-Noise ratio (HNR) is also obtained via Praat (9). Second, the autoregressive power spectral density (PSD) of the noise is obtained via Burg's method with respect to voice quality and glottal phase. Finally, the probability density function (PDF) of the noise is obtained and fitted with a t location-scale distribution, i.e., a Student's t-distribution, given by

$$St(x|\mu,\sigma,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left[\frac{\nu + \left(\frac{x-\mu}{\sigma}\right)^2}{\nu}\right]^{-\left(\frac{\nu+1}{2}\right)} \quad (1)$$

where μ is the location parameter, i.e., the mean, σ is the deviation parameter, i.e., the scale, and ν is the shape parameter. With $\nu = \infty$, the distribution equals a normal distribution. Decreasing ν make the main lobe lower and more narrow, and makes the tails heavier.

3. RESULTS

Figure 5 reports relative noise energy levels with respect to voice quality (breathy versus normal) and glottal phase (open versus closed). The medians of the noise energy levels are close to 0 dB for breathy voices, which would suggest that the breathy voices' noise energy is approximately equal to the energy of the voices' cyclic components. The energy distributions range from approximately -2 dB to 0.5 dB. One outlier exists at approximately -4.5 dB. As expected, noise energy levels of normal voices are lower on average. Their spread is larger, i.e., they range from approximately -10 dB to -1 dB. No differences between open and closed phases are observed for either breathy or normal voices. The difference in noise energy levels between breathy and normal voices is statistically significant.

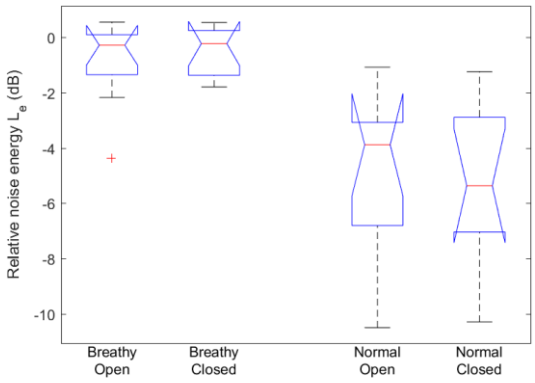

Figure 5 – Relative noise energy levels with respect to voice quality and cycle phase.

Figure 6 reports the HNR (dB) with respect to voice quality. For breathy voices, the HNR approximately ranges from 5 dB to 20 dB, with a median close to 14 dB. HNR is higher for normal voices, and ranges from approximately 18 dB to 29 dB. The median is close to 26 dB. The overlap between the distributions is small.

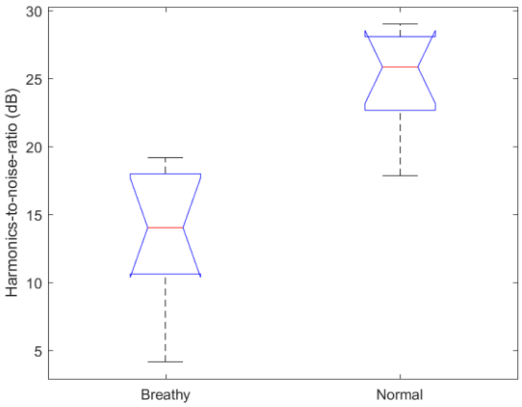

Figure 6 – Harmonics-to-Noise ratio with respect to cycle phase (9).

Figure 7 shows 10-th order autoregressive power spectral density estimates of the noise obtained via Burg's method and averaged over speakers (solid lines) and linear regression models thereof (dotted lines) with respect to voice quality (10). No relevant difference between the estimates is observed at low frequencies. The intercepts of the two regression lines are approximately equal (-97.5 dB/Hz and -97.7 dB/Hz). The difference in PSD estimates between breathy and normal voices increase with frequency. In particular, the slope of the regression lines is $-0.946 \cdot 10^{-3}\ dB/Hz^2$ for breathy voices, and $-1.461 \cdot 10^{-3}\ dB/Hz^2$ for normal voices.

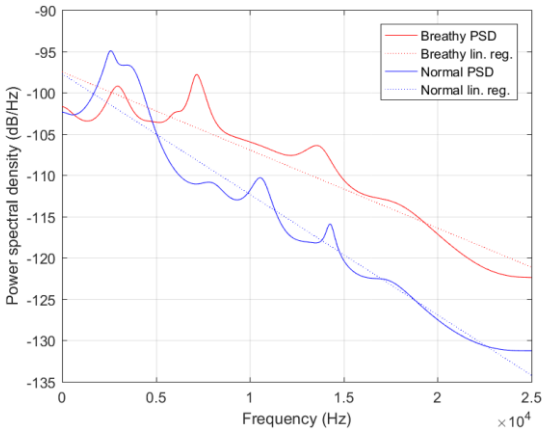

Figure 7 – Power spectral density (PSD) of the noise with respect to voice quality, and linear regression models of the PSDs.

Figure 8 (a) shows PSD estimates for breathy voices, and Figure 6 (b) shows PSD estimates for normal voices. Red and blue solid lines show estimates of open and closed phases respectively, and the dotted lines show linear regression models thereof. No relevant differences are observed between the open and closed phases for either breathy or normal voices.

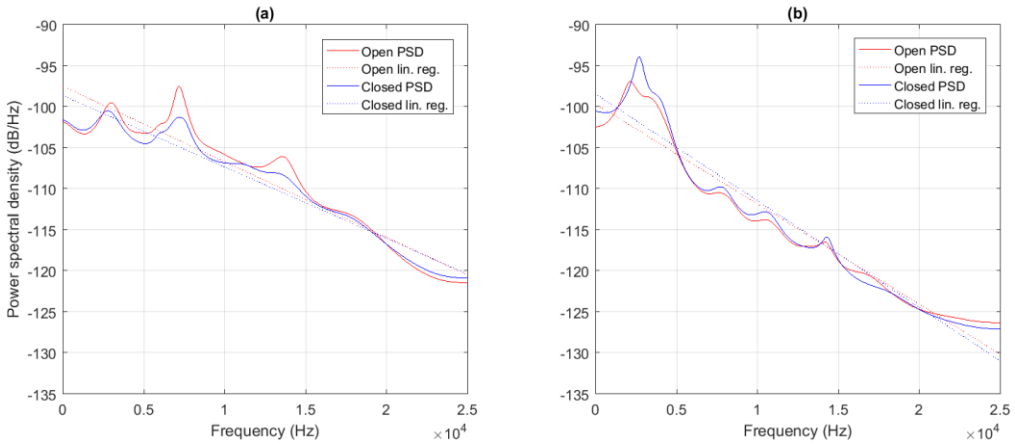

Figure 8 – Power spectral density (PSD) of the noise with respect to voice quality and glottal phase, and linear regression models of the PSDs. (a) Breathy voice quality, and (b) Normal voice quality.

Figure 9 shows a cumulative histogram (blue bars) of the model error time series values and a parametric fit (red lines). The solid line reports the overall fit, and the dotted lines report 95% confidence intervals. A t location-scale distribution is favored for modelling the noise signal over a Normal distribution, because of a better log likelihood. Table 1 summarizes the parameters of the fitted distributions.

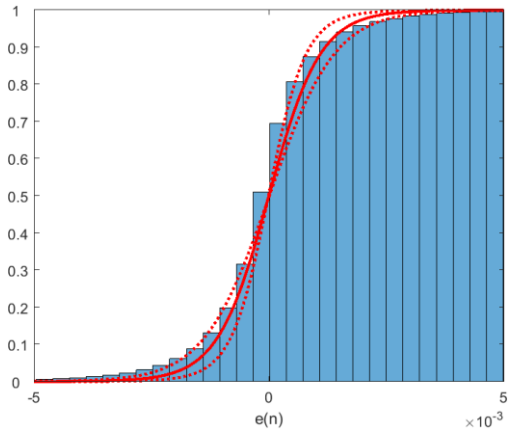

Figure 9 – Cumulative histogram (blue bars) of the noise, and its fitted t location-scale distribution (red lines). Shown are the average cumulative distribution fit (solid line), and 95% confidence intervals (dotted lines).

Table 1 – Minima, medians, and maxima of t location-scale distribution parameters of the model error signal time series' probability density distributions. Parameters are obtained subject-wise.

Parameter	Minimum	Median	Maximum
Location μ	$-2.091 \cdot 10^{-5}$	$-9.957 \cdot 10^{-7}$	$7.804 \cdot 10^{-5}$
Scale σ	$1.655 \cdot 10^{-4}$	$6.670 \cdot 10^{-4}$	$2.012 \cdot 10^{-3}$
Shape ν	3.165	6.469	12.63

4. DICUSSION AND CONCLUSION

A method for the segregation of additive noise from voiced speech signals produced by pathological and normal subjects is presented. The additive noise is described with respect to voice quality and glottal phase by means of energy levels, power spectral densities, and probability density functions. The results suggest updates of the existing modeling framework. First, spectral slopes that decay linearly with increasing frequencies are observed. The difference between breathy and normal voices is not only a difference in noise energy, but also in spectral slope. In particular, noise is boosted at high frequencies in breathy voices as compared to normal voices. No evidence in favor of a pulsatile additive noise component is observed. A t location-scale distribution is found to better fit the observed noise data than a normal distribution. Finally, the conventional Harmonics-to-Noise ratio (HNR) enables distinguishing between breathy and normal voices. The method that is presented here may enable analysis of additive noise in the presence of strong modulations of the vocal frequency and vocal cycle amplitude, which are frequent in pathological voices.

ACKNOWLEDGEMENTS

This work was supported by the Austrian Science Fund (FWF): KLI 722-B30.

REFERENCES

1. Dejonckere PH, Bradley P, Clemente P, Cornut G, Crevier-Buchman L, Friedrich G, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Eur Arch Otorhinolaryngol. 2001;258(2):77–82.
2. Fraj S, Schoentgen J, Grenez F. Development and perceptual assessment of a synthesizer of disordered voices. J Acoust Soc Am. 2012;132(4):2603–15.
3. Aichinger P, Roesner I, Leonhard M, Denk-Linnert D, Bigenzahn W, Schneider-Stickler B. A database of laryngeal high-speed videos with simultaneous high-quality audio recordings of pathological and

non-pathological voices. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC). 2016. p. 767–70.
4. Lohscheller J, Eysholdt U, Toy H, Dollinger M. Phonovibrography: Mapping High-Speed Movies of Vocal Fold Vibrations Into 2-D Diagrams for Visualizing and Analyzing the Underlying Laryngeal Dynamics. IEEE Trans Med Imaging. 2008;27(3):300–9.
5. Aichinger P, Hagmuller M, Schneider-Stickler B, Schoentgen J, Pernkopf F. Tracking of Multiple Fundamental Frequencies in Diplophonic Voices. IEEE/ACM Trans Audio, Speech, Lang Process. 2018;26(2):330–41.
6. Aichinger P, Pernkopf F, Schoentgen J. Detection of extra pulses in synthesized glottal area waveforms of dysphonic voices. Biomed Signal Proces. Elsevier Ltd; 2019;50:158–67.
7. Byrd RH, Gilbert JC, Nocedal J. A trust region method based on interior point techniques for nonlinear programming. Math Program Ser B. 2000;89(1):149–85.
8. Waltz RA, Morales JL, Nocedal J, Orban D. An interior algorithm for nonlinear optimization that combines line search and trust region steps. Math Program. 2006;107(3):391–408.
9. Boersma P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceedings of the Institute of Phonetic Sciences. University of Amsterdam; 1993. p. 97–110.
10. Kay SM. Modern Spectral Estimation: Theroy and Application. Modern Spectral Estimation: Theory and Application. Upper Saddle River, New Jersey: Prentice-Hall; 1988.