

# Robust Techniques for Pre- and Post-Surgical Voice Analysis

*Claudia Manfredi and Giorgio Peretti\**

\*Department of Electronics and Telecommunications, Univ. of Florence, Florence, ITALY

[manfredi@det.unifi.it](mailto:manfredi@det.unifi.it)

\*Clinic of Otolaryngology, Civil Brescia Hospital, Brescia, ITALY

[g.peretti@tin.it](mailto:g.peretti@tin.it)

## Abstract

Objective measure and tracking of the most relevant voice parameters is obtained for voice signals coming from patients that underwent thyroplasty implant. Due to the strong noise component and high non-stationarity of the pre-surgical signal, robust methods are proposed, capable to recover the fundamental frequency, tracking formants, and quantify the degree of hoarseness as well as the patient's functional recovery in an objective way. Thanks to its high-resolution properties, autoregressive parametric modelling is considered, with modifications required for the present application. The method is applied to sustained /a/ vowels, recorded from patients suffering from unilateral vocal cord paralysis. Pre- and post-surgical parameters are evaluated, that allow the physician quantifying the effectiveness of the Montgomery thyroplasty implant.

**Key words:** voice analysis, parametric modelling, pitch, noise, formants, vocal cord paralysis

## 1. Introduction

Any abnormality of the larynx that affects the vibration pattern of the vocal folds and the audible quality of the speech will be evident in the glottal waveform. There are a number of different causes of unilateral vocal cord paralysis. The most common cause is non-laryngeal cancer, which includes neoplasm of the head, neck, chest, and skull base. Neuritis associated with upper respiratory infection, syphilis, or other infectious sources may cause nerve dysfunction. General medical conditions such as diabetes mellitus may cause an isolated neuropathy, giving rise to vocal paralysis. Lesions of the vagal nerve occurring higher in the brain and may present with multiple cranial nerve abnormalities.

The Montgomery thyroplasty is a one-piece plastic prosthesis from a medical-grade polymer material, which is inserted by way of an incision made in a natural fold of the neck. Once inserted, the implant pushes the paralysed vocal cord to the midline so that the opposite functioning cord can contact it to phonate and close the airway when needed to prevent aspiration and produce an effective cough. The procedure helps to

re-establish the mucosal wave in the paralysed vocal fold. By approximating the vocal membranes, normal anatomic position is re-established and the cords are able to produce sound [8].

Due to high signal variability and strong noise, robust methods are required, capable to recover the speech fundamental frequency, quantify the degree of hoarseness as well as the post-surgical functional recovery, and allowing reliable formant tracking in an objective way. Several approaches exist to these problems, either in the frequency domain or in the time domain, but usually require a long time window for analysis. To overcome this problem, a variable window length is proposed, tailored to the varying speech characteristics. Fundamental frequency is evaluated by means of a two-step procedure, based on pitch estimation methods, revisited in order to enhance robustness to noise. Autoregressive PSD is considered, in order to recover formants' position [4], [5], [6]. Moreover, an adaptive version of the Normalised Noise Energy index [1] is proposed [3], with the aim of tracking the noise energy evolution within the utterance. This is indicative of the effectiveness of the adopted surgical technique, as well as of the post-surgical functional recovering. Finally, a quantitative SNR measure is defined, as an aid for surgical and rehabilitation effectiveness evaluation. The results show the good performance of the method, also for highly degraded voices.

## 2. Fundamental frequency estimation

Fundamental frequency  $F_0$  is estimated by means of a two step procedure. The first step is required in order to make the procedure completely automatic. The window length is chosen as  $M=3F_s/f_{\min}$ , where  $f_{\min}$  is the minimum allowed  $F_0$  value for the signal under consideration (50Hz, corresponding to very low male pitch). The following procedure is implemented on subsequent non-overlapped data frames [4], [6], [7], [8]:

- Singular Value Decomposition (SVD) is performed on a properly organised data matrix  $A$  of dimension  $(2(M-L) \times L)$ , where:  $M$ =window length,  $L$ =maximum allowed order for the filter ( $L=F_s+4$  in this work). This gives the optimal signal subspace dimension,  $p$ ;

- AutoRegressive, AR(p) parameters are estimated, for the all-pole vocal tract inverse filter [5], which is applied to the signal, thus obtaining the residual sequence (band-pass filtered in the range 50Hz-1kHz);
- The maximum of the autocorrelation sequence (AS) of the residuals is evaluated in the frequency range of interest (50-400Hz). The pitch value is then given by  $F_0 = F_s / \tau$ , where  $\tau = \arg(\max(\text{AS}))$ .

This procedure gives the range of variation for  $F_0$  during the pronounced utterance:  $F_1 \leq F_0 \leq F_h$ ,  $F_1$  and  $F_h$  being the minimum and the maximum estimated pitch value respectively.

In the second step,  $F_0$  is adaptively estimated in the frequency range  $[F_1, F_h]$ , on data frames overlapped for  $3/4$  of length, according to the following steps:

- On each varying time window, the Mexican Hat wavelet transform is applied (MHW) [2], [4]. Its scale parameter  $s$  is allowed to vary in the range 1-80 (achieved experimentally). The time parameter  $k$  varies in the allowed range (linked to the variable window length). This gives a coefficient matrix,  $\text{MHW}(k, s)$ ;
- From  $\text{MHW}(k, s)$ , the optimum scale value,  $\hat{s}$ , is selected as the one corresponding to the maximum entry: this in fact represents the best fitting of the wavelet to data.
- The Average Magnitude Difference Function (AMDF) is applied to  $\text{MHW}(k, \hat{s})$ , thus obtaining the estimate of  $F_0$  as:  $F_0 = F_s / \eta_{\min}$ , where  $\eta_{\min}$  is the AMDF minimum.

The choice of the AMDF instead of the autocorrelation sequence (AS) is due to the non-stationarity and amplitude modulation of the signals under study. These aspects were shown to often cause misestimating the true signal periodicity with the AS [6].

### 3. Noise estimation

The Normalised Noise Energy (NNE) acoustic measure [1] is a measure of the dysphonic component of the voice spectrum related to the total signal energy.

Given the speech signal  $x(n) = s(n) + w(n)$ , where  $s(n)$  is the periodic component and  $w(n)$  is the additive noise component, let  $X(k)$ ,  $S(k)$  and  $W(k)$  be the Discrete Fourier Transform (DFT) of  $x(n)$ ,  $s(n)$  and  $w(n)$  respectively. The Adaptive NNE (ANNE) is defined as:

$$\text{ANNE}(k) = 10 \log \left[ \frac{\sum_{m=N_L}^{N_H} |\tilde{W}_m(k)|^2}{\sum_{m=N_L}^{N_H} |X_m(k)|^2} \right], \quad k = N_L, \dots, N_H \quad (1)$$

with:  $N_L = \lceil N f_L T \rceil$ ,  $N_H = \lceil N f_H T \rceil$ ,  $N$  = number of DFT points,  $L$  = number of frames in the analysis interval,  $F_L$  and  $F_H$  = lowest and highest frequencies of the frequency band of interest, respectively.  $|\tilde{W}_m(k)|^2$  is an

estimate of the unknown noise energy  $|W_m(k)|^2$ ,  $|X_m(k)|^2$  is the signal energy and  $T$  is the sampling period. In the harmonic dip regions  $D_i$  (i.e. where the harmonics have no component)  $|\tilde{W}_m(k)|^2$  is given by the signal energy  $|X_m(k)|^2$ , while in the harmonics peak regions  $P_i$  it can be obtained by interpolating between the values of  $|\tilde{W}_m(k)|^2$  in the dip regions  $D_i$  and  $D_{i+1}$  on both sides of the peak region  $P_i$ . The width of peak and dip regions is related to the choice of the window function. Here, as in [1], a Hamming window is considered, whose approximate bandwidth in terms of  $k$  is  $2N/M$ ,  $M$  being the number of points in the analysed window. Hence:  $D_i = \{k \mid k_{i-1} + 2N/M < k < k_i - 2N/M\}$ ,  $P_i = \{k \mid k_i - 2N/M \leq k \leq k_i + 2N/M\}$ , where  $k_i$  and  $k_{i-1}$  are the  $i$ -th and  $(i-1)$ -th harmonic peak location respectively. Notice that the dip region  $D_i$  may disappear for a fixed value of  $M$ , as the  $i$ -th and  $(i-1)$ -th harmonic peak locations get closer to each other. The method proposed here allows a period-to-period adaptive choice of the minimum time-window length  $M$ , which guarantees non-empty dip regions, for fixed  $N$  and  $F_s$ , also in the case of a pitch period varying within the same utterance. In [4] it was shown that  $M$  given by:

$$M(F_0) \geq \frac{4N}{F_0 \frac{N}{F_s} - (d+1)} \quad (2)$$

is optimal as far as tracking of signal spectral characteristics is concerned. The ANNE is thus evaluated according to the following steps:

- Select the sampling frequency  $F_s$ , the number  $N$  of DFT points and the number  $d$  of points in the dip regions (Here:  $F_s = 50\text{kHz}$ , according to the data acquisition device,  $N = 16384$  and  $d = 10$ ). This gives the adaptive time window length  $M$ , tailored to  $F_0$  variations (eq.4);
- Multiply each data window by a Hamming window and evaluate the power spectrum on that frame;
- Evaluate the ANNE on  $3/4$  partially overlapped data frames.

Notice that the lower the ANNE, the lower the voice hoarseness.

### 4. Formant estimation

Parametric formant estimation relies on a vocal tract model made up by an interconnected series of  $p$  cylindrical coaxial lossless cavities of different length and diameter. The resonances (formants) can be recovered from the maxima of the power spectral density (PSD), given by:

$$\text{PSD}(f) = \frac{T}{\left| 1 + \sum_{k=1}^p a_k e^{-j2\pi k f T} \right|^2} \quad (3)$$

where  $T$  is the sampling period and  $a_i$ ,  $i = 1, \dots, p$ , are the coefficient of the AR model of order  $p$  that describes the vocal tract [5]. One of the main advantages of

parametric spectral analysis consists in its high-resolution capability, as the model extrapolates data outside the analysed window. However, AR spectral estimators are very sensitive to order selection: in case of overestimated model order  $p$ , formant splitting may occur, while underestimation smoothes the spectrum and causes misallocation of spectral peaks. Many criteria have been defined for finding the best model order  $p$ , including both the estimated variance  $\sigma^2$  and the model complexity  $p$  in one statistics. Such criteria are characterised by loss functions for which a minimum can be achieved. However, such criteria were shown to be almost unreliable for short data frames, due to long-term convergence properties. In this paper, the relation  $p \approx F_s$  was found the best one for obtaining a sufficiently detailed spectrum.

## 5. Quality measure

The proposed objective index for measuring voice quality enhancement is evaluated as the ratio (in dB) between the noisy signal energy and that of the amount of noise disregarded after the surgical intervention, and is defined as:

$$SNR_y = 10 \log_{10} \frac{\sum_{n=1}^M y^2(n)}{\sum_{n=1}^M (y(n) - y_{\text{post}}(n))^2} \quad (4)$$

Where:  $y(n)$ ,  $y_{\text{post}}(n)$ =signal sample at time  $n$ , before and after surgical intervention, respectively, and  $M$ =data length. Hence, low  $SNR_y$  values correspond to high voice quality enhancement.

## 6. Experimental results

The new procedure was successfully applied to simulated signals, with different SNR and jitter values, and to real pathological voices [4], [6], [7].

In this work, pre- and post-surgical comparison of  $F_0$ , ANNE,  $SNR_y$  and formants is considered, in order to objectively evaluate the effectiveness of Montgomery thyroplasty in patients affected by unilateral vocal cord paralysis. 25 patients (10 male and 15 female) with glottis insufficiency of varying aetiologies underwent Montgomery I thyroplasty at the E.N.T. Department of the University of Brescia, Italy. All subjects underwent a battery of clinical-instrumental tests pre- and postoperatively as well as 2, 6 and 12 months after surgery. Videolaryngostroboscopy showed that the glottis closed completely in most cases and objective acoustic analysis parameters were significantly improved after surgery.

Tracking of  $F_0$ , ANNE and formants was obtained for each patient, for pre- and post-surgical voice recordings. The sampling frequency was 50kHz. The following figures, Figs.1-4, are relative to a female patient. Voice signal,  $F_0$ , ANNE are plotted, before (Fig.1) and after

(Fig.2) the thyroplastic intervention. Before surgery, voice hoarseness was extremely high, as can be argued from the signal plot (upper plot in Fig.1), which shows irregular, noise-like and low-energy characteristics. As a consequence,  $F_0$  was hardly recognised and shows large variations (middle plot), ranging from about 100Hz up to 300Hz. Both the first (circles) and the second (crosses)  $F_0$  estimation steps were applied. ANNE estimation (lower plot) confirms the strong noise component present in this signal, with a mean-value around -3 dB.

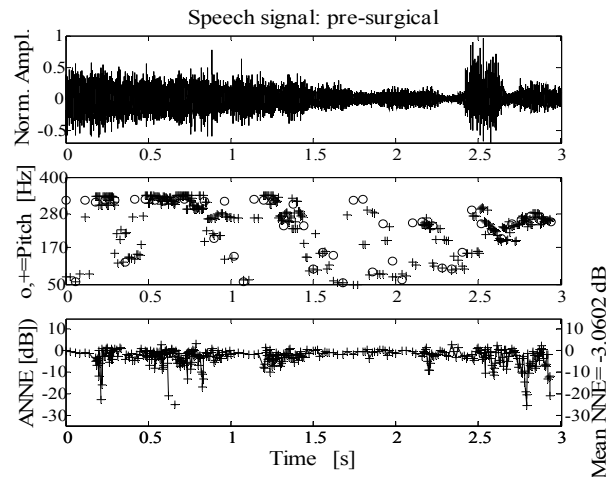


Figure 1: Pre-surgical voice analysis. Upper: voice signal; middle: two-step  $F_0$  estimation; lower: ANNE.

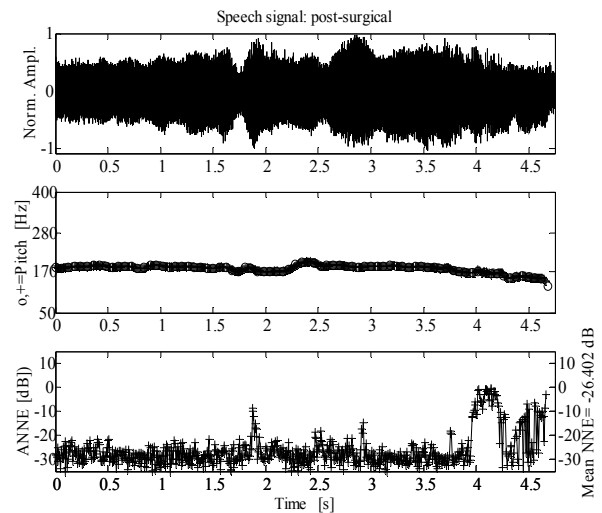


Figure 2: Post-surgical voice analysis. Upper: voice signal; middle: two-step  $F_0$  estimation; lower: ANNE.

From Fig.2, voice improvement is evident, as both the signal and  $F_0$  exhibit a more regular behaviour. Voice strength results enhanced (upper plot) and  $F_0$  has stabilised around 180 Hz (middle plot). The ANNE values are largely decreased, with a mean value of about -26dB (lower plot).

Notice that pre- and post-surgical changes in  $F_0$  were observed in almost all cases. This could be due both to

the noisy-like signal before surgery and to the Montgomery prostheses, which might alter the natural voice parameters.

Fig.3 compares formant tracking obtained before (stars) and after (circles) surgery. Thanks to recovered phonation, formants are now clearly visible and show an almost regular behaviour.

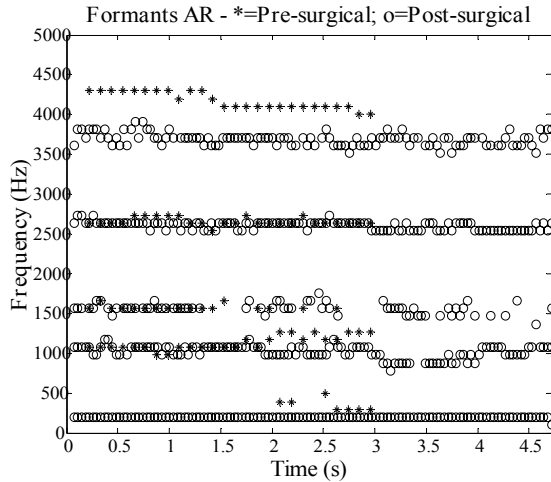


Figure 3: AR formants estimation (model order  $p=50$ ) before (\*) and after (o) thyroplasty implant.

Concerning  $SNR_y$  values, in all cases the values were found in the range  $-2\text{dB}/-11\text{dB}$ , showing that the quality of voice was always enhanced. Specifically, the lowest  $SNR_y$  values ( $<-7\text{dB}$ ) correspond to highly degraded voices, with energy below or near  $0\text{ dB}$  in the  $0\text{-}1000\text{Hz}$  frequency range before surgery. Higher values were found to be relative to less dysphonic voices.

Fig.4 shows the increase of PSD after the surgical intervention, mainly in the low-frequency region, where vocal fold paralysis prevented phonation. Notice the low SNR value (about  $-9\text{ dB}$ ), which confirms the effectiveness of the thyroplasty implant.

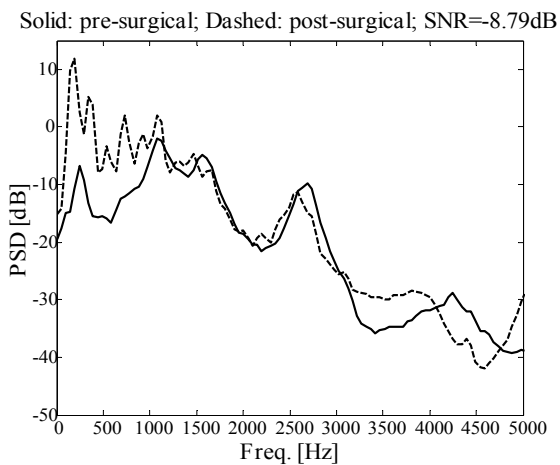


Figure 4: AR PSD before (solid line) and after (dashed line) the surgical intervention.  $SNR=-8.79\text{ dB}$ .

## 7. Final remarks

This paper presents first results concerning tracking of the most relevant voice parameters for a class of pathological hoarse voices. Due to the high signal variability, robust methods are proposed, based on autoregressive parametric modelling, capable to recover the fundamental frequency, quantify the degree of hoarseness and estimate formants in an objective way. The method is applied to sustained /a/ vowels, recorded from patients suffering from unilateral vocal cord paralysis. Pre- and post-surgical parameters are evaluated, that allow the physician quantifying the effectiveness of the Montgomery thyroplasty implant. Future work will be devoted to compare the results with those obtained with the MDVP software tool, as well as to relate them with subjective scales. Moreover, observed changes in fundamental frequency and formants values have to be further exploited, also by means of AR based spectrograms, in relation to the adopted surgical technique.

## 8. References

- [1] Kasuya,H., Ogawa,S., Mashima,K. and Ebihara,S., "Normalised noise energy as an acoustic measure to evaluate pathologic voice", *J. Acoust. Soc. Am.*, 80(5): 1329-1334, 1986.
- [2] Kadambe,S. and Bourdeaux-Bartels,G.F., "Application of the Wavelet transform for pitch detection of speech signals", *IEEE Trans. Inf. Theory*, 38(2): 917-924, 1992.
- [3] Manfredi,C., "Adaptive noise energy estimation in pathological speech signals", *IEEE Transactions on Biomedical Engineering*, 47(11): 1538-1542, 2000.
- [4] Deller,J.R., Proakis,J.G. and Hansen, J.H.L., *Discrete-time Processing of Speech Signals*, Maxwell McMillan, New York, 1993.
- [5] Manfredi,C., D'Aniello,M., Brusaglioni,P. and Ismaelli,A., "A comparative analysis of fundamental frequency estimation methods with application to pathological voices", *Medical Engineering and Physics*, 22(2): 135-147, 2000.
- [6] Fort,A., Ismaelli,A., Manfredi,C. and Brusaglioni,P., "Parametric and non-parametric estimation of speech formants: application to infant cry", *Medical Engineering and Physics*, 18(8): 677-691, 1996.
- [7] Fort,A., Manfredi,C. and Rocchi,S., "Adaptive SVD-based AR model order determination for time-frequency analysis of Doppler ultrasound signals", *Ultrasound in Medicine and Biology*, 21(6): 793-805, 1995.
- [8] Peretti,G., Provenzano,L., Piazza,C., Giudice,M. and Antonelli,A.R., "Functional Results After Type I Thyroplasty with Montgomery Prosthesis", *Acta Otorhinolaryngol Ita*, 21(3): 156-162, 2001.