VOQUAL'03, Geneva, August 27-29, 2003

Improvement in hoarse voice denoising for real-time DSP implementation

Claudia Manfredi, Fabrizio Dori, Ernesto Iadanza

Department of Electronics and Telecommunications
Università degli Studi di Firenze, Via S. Marta 3, 50139 Firenze, Italy
E-mail: manfredi@det.unifi.it

Abstract

Voice hoarseness is mainly related to airflow turbulence in the vocal tract. It can be due to vocal fold paralysis, polyps, cordectomisation or other dysfunction, which alter regular speech production, and is commonly treated as a noise component in the speech signal. A denoising approach is proposed, based on low-order singular value decomposition (SVD) of matrices whose entries come from sampled speech data frames, properly organised. A prototype DSP board implementing the procedure was developed, by means of properly optimised C and Assembler code. Enhanced results are obtained with respect to a previous scheme, by introducing a normalization step on the signal amplitude dynamics. This allows increasing the output level, as well as reducing click-noise, both due to the algorithm structure and DSP implementation constraints. Objective quality indexes are proposed, showing the better results achieved with the proposed modifications.

1. Introduction

This paper deals with the problem of enhancing voice quality for people suffering from dysphonia. This can be due to vocal fold paralysis, polyps, cordectomisation or other dysfunction, which alter regular speech production and commonly cause more efforts to be used in speaking than for healthy people. The quality of speech signals is a measure, which reflects on the way the signal is perceived by listeners. It can be expressed in terms of how pleasant the signal sounds or how much effort is required on behalf of the listeners in order understand the message. Subjective measures are based on the opinion of a group of listeners on the quality of an utterance. The disadvantage of these tests is that there may be variations or biases among listeners. Moreover, they require significant time and personnel resources. Objective speech quality measures are reliable, easy to implement and have been shown to be good predictors of subjective quality [7], [17]. The main goal of the system presented here is to realise a mobile hardware/software system for real-time voice denoising, to obtain a more intelligible speech. The method is based on the singular value decomposition (SVD) of matrices whose entries come from sampled speech data frames, properly organised [1]. The SVD is a powerful tool that allows separating the signal and the noise component in noise corrupted signals. Due to its robustness against noise, SVD is widely used for speech enhancement, mainly to improve the performance of speech communication systems in a noisy environment [2], [3], [4]. For the present application, a fixed two-dimensional signal subspace dimension was found sufficient for data filtering, thus allowing real-time implementation. Objective quality measures (PSD ratios, SNR) are defined end evaluated, in order to assess enhancement of voice and compare results. The reconstructed signal shows that the spectral characteristics of the original signal are preserved, with low PSD ratio values in the low frequency region (<4kHz) and high values above, where especially noise tends to make the harmonic structure unclear. The proposed approach was implemented on a DSP board, by means of properly optimised C and Assembler code. Thus, a simple portable device could be realised, as an aid for dysphonic speakers. It could be of help for diminishing effort in speaking, which is closely related to social problems due to awkwardness of voice. A high quality microphone would collect low level voice and produce clean and sufficiently loud voice at the output of the device, connected with loudspeakers. For portability, it would also be easy to use, light and small. A prototype is under study.

2. Denoising with SVD

The SVD is a numerically reliable and robust means for estimating the space of clean data (signal subspace) from the white noise corrupted data, and is thus particularly suited for speech denoising. It performs the factorisation: $A=U\Sigma V^H$, H denoting transpose-conjugate, for a matrix A, generally complex-valued and non-square. Matrix Σ is block-diagonal, with the (1,1) block given by: $\Sigma_p=\text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p)$. The singular values σ_i display the distance of matrix A from low-rank matrices and together with the singular vectors they can be used to construct optimal low-rank approximants, A_p, where p is the size of the low-rank approximation [1],[5],[6]. This low-rank approximation considerably improves the quality of the parameter estimates, because the singular vectors corresponding to the cluster of small singular values (i.e. the noise subspace) will no longer contribute to the solution, thus removing a major source of sensitivity to noise.

The SVD method was applied on real-valued, 2(N-L)xL A matrices (hence, in this case, $A=U\Sigma V^T$, T denoting transpose), where L is the maximum allowed filter order and N is the data frame length. Here $L=F_s$ (in kHz), where F_s is the sampling frequency [7]. Notice that L is chosen such that L>p. In fact, overmodeling combined with the minimum-norm choice turns out to be an effective way to overcome parameter sensitivity. A-matrix structure is Toepltiz-like, and arises from the classical forward-backward approach to the estimation of linear prediction (LP) polynomial coefficients. Its entries are obtained from subsequent N-points data frames, with $N=3F_s/F_0$, F_0 being a reasonable minimum value of the fundamental frequency for the signal under consideration (as data are relative to adult male voices F_0=50Hz). Motivations for this choice of N can be found in [8].

The following steps are then performed on A:

- Compute the SVD of $A = U\Sigma V^T = \sum_{k=1}^{r} \sigma_k u_k v_k^T$, $r=\min(L,2(N-L))\geq p$, u_k and v_k are respectively the left and right singular vectors associated with the eigenvalue σ_k.
- The p-rank approximation of A is obtained by retaining the p dominant singular values and the corresponding singular vectors, i.e. $\hat{A}_p = \sum_{k=1}^{p} s_k u_k v_k^T$.
  Hence, $\hat{A}_p$ corresponds to Σ as far as the first p eigenvalues are concerned, and is zero elsewhere.
- From $\hat{A}_p$, the filtered signal frame is reconstructed. The subsequent N-points speech frame is analysed. Filtered frames are put back together sequentially, appending the new frame to previously filtered frames.

Notice that SVD requires selecting the "size" p of the signal subspace, i.e. the minimum number of eigenvectors spanning the clean data. This is commonly achieved by separating the largest singular values of Σ from the smallest ones by means of suitable thresholds. To this aim, variable and static thresholds were tested, giving $2\leq p\leq 6$ during the utterance [9]. As it was found that the higher the order p, the worse the filter, a fixed low-order filter was selected, corresponding to p=2.

Despite its simplicity, the SVD approach was found effective in increasing voice quality. Extensive simulations were performed and detailed results are reported in [9], [10]. This will be called the 'original' or 'standard' approach.

This paper aims at overcoming some limitations and drawbacks of the original method, by introducing small but effective changes in the software tool, that will be described in what follows. Such modifications will be addressed to as the 'enhanced' or 'improved' method. Moreover, in order to measure such improvements, some simple objective quality indexes will be introduced and evaluated.

Two residual problems were in fact found, regarding signal quality after the filtering process. The first one concerns the signal level, as the output level was significantly lower than the input one. This problem is apparently due to scale factors that are required in the filtering algorithm. In fact, they allow avoiding overflow problems, while preserving the high precision level requested by the specific application. This problem has been addressed downstream, introducing a normalization step of the signal amplitude dynamics in the improved method. We are also testing new implementations of filtering algorithm that include normalization blocks interposed between filtering blocks.

The second problem concerns appearing of a new undesired signal superimposed to the filtered signal, audible as a fast series of "clicks". Click-noise problem is intrinsic in this filtering chain, that is built starting from non-overlapping frames of the input signal. This choice comes from the strong requirement of saving resources, due to the notable dimension of matrices involved in computations. In this work, this problem is addressed with a linear interpolation across n samples of the filtered signal, centered on the last sample of each frame. Our tests have shown that n=5 (i.e., 2 samples before and 2 after the final sample of each frame) is a good choice for improving audible quality of the output file; more tests with n>5 do not show significant improvements. Interpolation allows signal smoothing with a simple operation, without requiring loading computational resources. At the same time, it is easy to upgrade to higher order function implementation, in order to achieve the best performance in terms of SNR and audible quality. At present, the development of the enhanced algorithm goes towards the simultaneous implementation of standard overlapping techniques and new methods for fast and light SVD calculation.

3. Quality measures

Extensive research has been carried out in developing both subjective and objective tests to ascertain quality, but few results are available as far as correlation among them is concerned [17]. In the following, some indexes are proposed, closely related to the signal characteristics. The subscript "non-filt" refers to the original signal, while "filt" refers to the SVD-filtered signal. The simplest one is:

$$PSD = 10\log_{10}\frac{PSD_{non-filt}}{PSD_{filt}} \qquad (1)$$

representing the ratio of the PSDs, evaluated on the whole frequency range;

$$PSD_{low} = 10 \log_{10} \frac{PSD_{non-filt}(f \le 4kHz)}{PSD_{filt}(f \le 4kHz)} \quad (2)$$

measures the ratio of the PSDs evaluated on the "harmonic" range, while

$$PSD_{high} = 10 \log_{10} \frac{PSD_{non-filt}(f \ge 4kHz)}{PSD_{filt}(f \ge 4kHz)} \quad (3)$$

is the ratio of the PSDs, evaluated on the "noise" range. A good denoising procedure should give PSD and PSD_{low} values around zero (no loss of power), but high PSD_{high} values (loss of power due to noise). Finally,

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

where: y(n) = noisy signal sample at time n, $y_{filt}(n)$ = filtered signal sample at time n. SNR is thus the ratio between the signal energy and that of the measured noise. Hence, low SNR values correspond to strong filtering.

Notice that PSD_{low} and SNR have good correlates with NHR [17] and the GIRBAS scale, while being simple and reliable at a very low computational cost. Hence, they could be of further help to the physician, as objective measures for assessing enhancement of voice. This point will be further exploited in future work.

4. Experimental results

The denoising procedure was applied here to real data. These concern hoarse pathological voices, coming from adult male subjects that underwent partial cordectomisation, due to T1A glottis cancer. Patients were asked to pronounce the Italian word /aiuole/ (flowerbeds), which is composed of the five principal vowels /a/, /e/, /i/, /o/, /u/. This choice is due to the clinical interest in evaluating the effort in speaking made by patients, for surgical and rehabilitation purposes. Seven laser and nine lancet cordectomised male subjects were analysed.

The enhanced SVD filtering procedure was compared to the original one, and the two methods were tested by means of the quality indexes described in sect.3. In all cases, the new approach was capable to better enhance the quality of voice. Figs. 1-3 show the results relative to one subject (lancet operated). The sampling frequency is F_s=25kHz. Fig.1 plots the PSD evaluated for the non-filtered signal (solid line) and for the signal filtered without and with the proposed improvements (dashed line, upper and lower plots, respectively). The original approach lowers the PSD on the whole frequency range (PSD=0.56 dB), and especially on the high frequency region (PSD_{high}=20.76 dB), but has a small negative effect on the low one (PSD_{low}=0.54). As previously noticed, this causes some lowering of the output signal level. Strong denoising is achieved, as can be inferred from the low SNR value (SNR=7.42 dB). As for the improved approach, better values are obtained for the signal power on the whole frequency range (PSD=0.02 dB), and especially on the low frequency range (PSD_{low}=-0.004 dB). This corresponds to a good voice level at the output of the filtering chain. As a side effect, lower PSD is found on the high frequency region (PSD_{high}=14.6 dB), and correspondingly a higher SNR value (SNR=16.4 dB). These results suggest that a selective procedure should be defined for the normalisation step, in order to avoid enhancing the filtered signal on undesired frequency ranges.

Figure 2 shows the spectrogram of the unprocessed signal (upper plot), as compared to those obtained from the standard SVD filtering (middle plot) and the enhanced one (lower plot). For clearness, the frequency range is limited to a maximum of 6 kHz. Both the middle and the lower plots confirm the good denoising properties of the proposed procedures, as the noise level is largely reduced (especially in the middle plot), above 4 kHz. However, the middle plot clearly shows the presence of click noise, which appears as almost regularly spaced vertical lines of rather high intensity. The lower plot shows that the new approach allows reducing this side effect, that has almost disappeared. Moreover, harmonics have been enhanced and often recovered, as their intensity colour clearly shows.

Finally, fig.3 reports formant trajectory evolution for the unprocessed signal (marked with "o"), as compared to the previous denoising procedure (symbol "*") and to the enhanced one (symbol "+"). Formant trajectory tracking is obtained by means of the Autoregressive (AR) PSD estimation method, with a model order p=25, corresponding to the signal sampling frequency. This choice, as described in literature [6], [7], is in fact a good compromise between parsimony and good resolution. As already said, denoising with the proposed SVD approach preserves the temporal and spectral characteristics of the original signal, thus providing a filtered voice of better quality, without distorting effects. The normalisation step introduced in the improved method results in sharper peaks in the PSD. This allows better formant tracking with the AR parametric approach, as shown in fig.3, especially in the low frequency region: more peaks, represented by the '+' sign, are clearly found below 2500 Hz. Hence, a more accurate peak picking is obtained.

5. Software implementation

The software development tool is the Code Composer Studio, which integrates a C compiler for source file translation into Assembler, a C linker for file and libraries linking, and the DSP/BIOS firmware for implementing a basic kernel with run-time services [11]. The SVD algorithm is implemented by means of a two-step procedure: first, the data matrix A is bi-diagonalised applying a sequence of Householder reflections; second, A is made diagonal using a modified QR algorithm [12-15]. For each data frame, four C routines (functions) are required: the first two build data matrices, the third one performs computations and the last one collects filtered data for the output frame reconstruction [16].

Data stream (300 samples) and singular values (30 samples) are stored in the on-chip memory, while matrices U and V (vectorised) are stored on the external SBSRAM memory, integrated on board. Assembler code is used for heaviest computations, i.e. those concerning SVD factorisation. Pipelining is used for loops.

The filtered signal is reconstructed by performing the product $\hat{A}_p$ of matrices U, Σ_p and V, with p=2, and by extracting the output stream from it. To this aim, as the location of these data inside $\hat{A}_p$ is known, only 300 products are needed. Computations are further reduced taking into account that only two singular values of Σ_p are non-zero.

6. Hardware implementation

Simulation studies performed under Matlab environment (rel. 5.2) have shown the algorithm is not suitable for real-time implementation on a personal computer. This is due to two main constraints:

1) high computational burden for matrix A factorisation and singular values evaluation;

2) large frame size to be processed in real-time.

It is worth to notice that the speech frame length depends on two parameters: the pitch value and the sampling frequency. The analysis performed on actual voice samples has shown that 100 Hz is a reasonable pitch value. As a consequence, the sampling frequency was set to 12kHz, as it represents a reasonable compromise between signal quality and frame length (300 data elements).

The criteria adopted to implement the hardware platform are:

- High processing performance.
- Ease and flexibility of the programming environment.
- Low power consumption.
- Low cost.

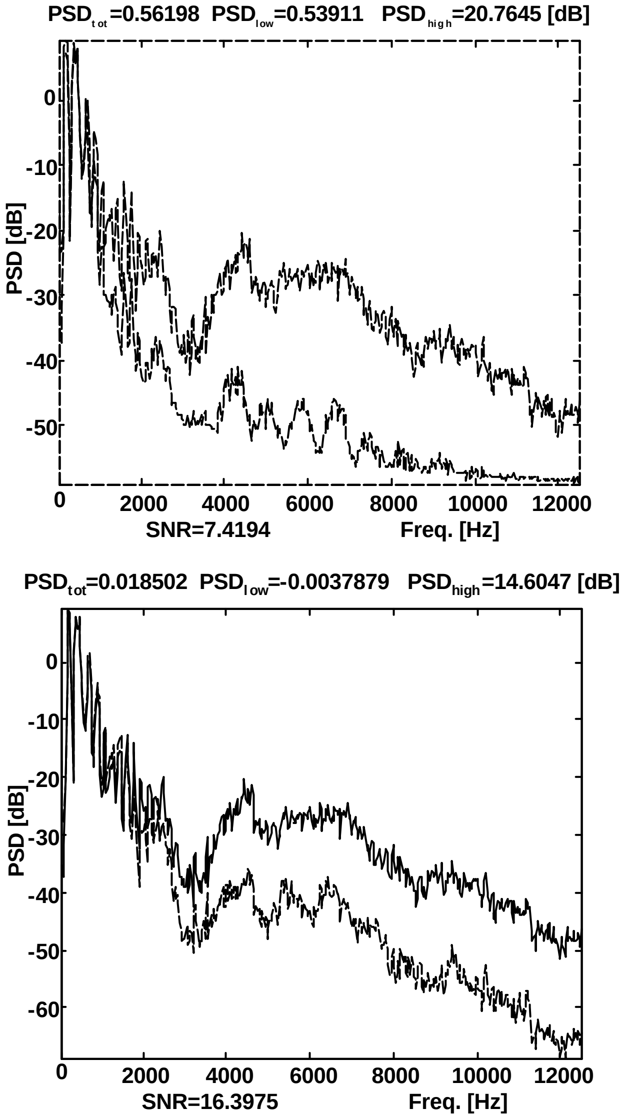

Figure 1 – Comparison of PSD plots before and after filtering. Upper: previous result; Lower: enhanced result

The board is supplied with analog front-end, capable to accept the audio signal as input and to furnish the output processed signal at the output stereo jack. The DSP-based board allows to process signals in the 0-48kHz bandwidth.

The main components of the board are:

- A floating point C6711 of the TMS320C6000 DSP family, developed by Texas Instruments with advanced velocity VLIW (Very Long Instruction Word) architecture. Such architecture is particularly suitable for multifunctional and multichannel applications. Due to its limited memory, only high-access data are resident on the on-chip memory, while low-access data are stored on the peripheral

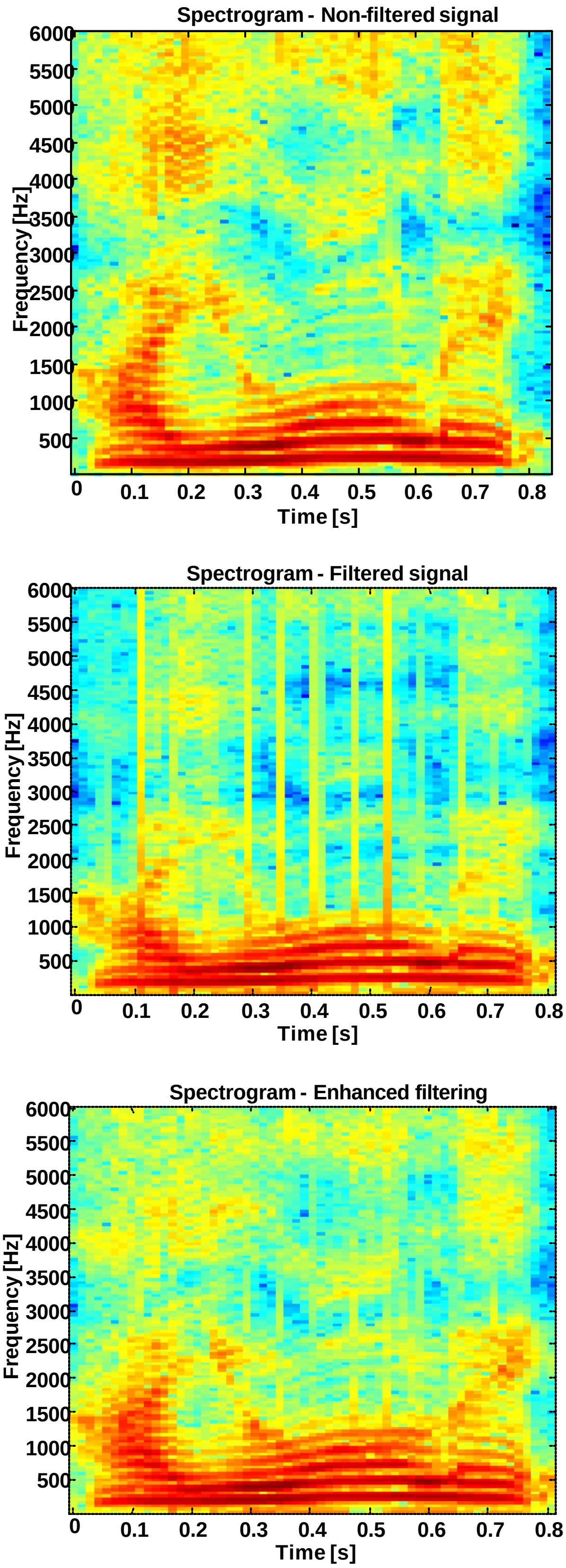

Figure 2 – Spectrogram of the signal before denoising (upper), after denoising (middle) and after enhanced denoising (lower).

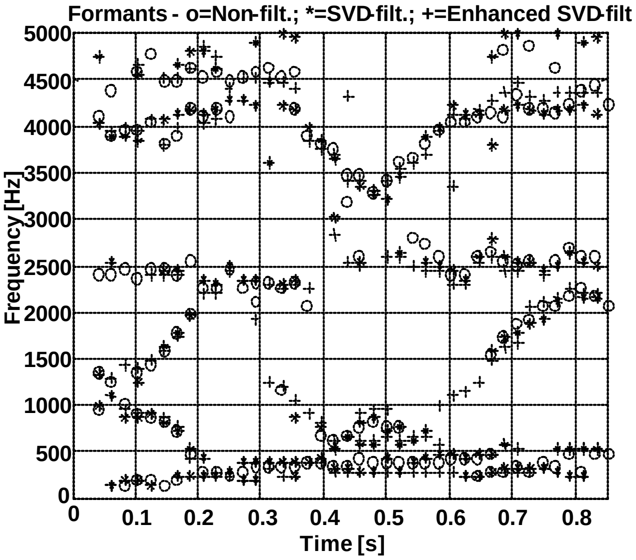

Figure 3 – Formant tracking by means of AR PSD. The AR model order is p=F_s (in kHz).

memory.

- An analog input interface. It detects the voice by a microphone and, through a pre-amplifier circuit, it feeds a differential amplifier directly connected to the A/D converter.
- An A/D converter. It is a sigma-delta (TLC320AD57-TI) serial converter, with improved performance with respect to conventional ones (flash or successive approximations).
- A D/A converter (PCM1725-TI). It is a low-cost stereo converter, made up by a 3rd order delta-sigma, a digital filter and an analog amplifier. It receives data through the McBSP gate and directly feeds the output stereo jack.
- An analog output interface made up of a low-pass circuit with 50kHz cut-off frequency.
- A DSP clock circuit, based on a phase-locked loop circuit (ICS501 - ICS), that reaches a maximum clock frequency of 160MHz.
- A SBSRAM memory (MT58L128L32DT7.5-Micron) with high-speed and low-power consumption. Such device supports 7.5 ns per cycle, i.e. 133 MHz clock frequency.
- A SRAM (IDT71V416L412 - IDT) high-speed 16 bit CMOS technology, made up of two parallel 16 bit memory benches, to allow simultaneous 32 bit data access.
- A flash memory (Am29LV040B-AMD) organised as 8x512 bit. It contains the boot code to be loaded on the DSP.

The developed hardware was tested with real data in order to reach the real-time processing requirements. Because the input data stream is made of 300 data samples, and the A/D converter operates at 12 kHz, the data at the DSP input changes with a T period given by: T=300/12000=25ms. Therefore, in order to perform real-time operations, the n-th frame available at time t, must be executed before the (t+T) instant, when a new (n+1)th frame is available at the DSP input.

In a first release, the execution time was 385 ms, with a non-optimised (high level) code. Using hand-coded assembler, the execution time reduces to 22.18ms, that perfectly agrees with real-time requirements.

7. Final remarks

A simple approach for enhancing voice quality in dysphonic subjects is proposed. The method applies SVD for data filtering, separating the clean signal from its noisy component. The denoised signal is reconstructed along the directions spanned by the principal eigenvectors of the signal subspace. For filtering purposes, the best choice was found that of picking only the two dominant eigenvalues, thus resulting in a low-cost procedure, suitable for on-line implementation on a DSP board. Real data coming from cordectomised adult male subjects were filtered with the proposed approach, with enhanced results in all cases.

With this approach, two main drawbacks were found. One concerns the output level of the signal, which resulted reduced after filtering. The second one consists in click-noise, that degrades the quality of the filtered voice, and that was introduced into the signal by processing the data on subsequent non-overlapped windows. A first attempt to solving both problems is described in this paper. The first problem was solved downstream, introducing a normalization step of the signal amplitude dynamics, while for the second one linear interpolation is performed across n samples of filtered signal, centered on final sample of each frame. Future work will concern testing new implementations of the filtering algorithm, that include normalization blocks interposed between filtering blocks. Moreover, normalisation should be performed on a selective basis, in order to disregard the spectral components that are mainly relative to noise.

Pre- and post-surgical voice quality has been exploited also by means of subjective indexes such as the GIRBAS scale, as well as the objective MDVP software tool. The results, which will be presented elsewhere, show the improvement in voice quality in almost all cases and are in accordance to those obtained with indexes proposed here.

8. References

[1] Rao B D, Arun K S., "Model based processing of signals: a state space approach", *Proc. IEEE*, vol.80, 1992, pp. 283-309.

[2] Asano F, Hayamizu S, Yamada T, Nakamura S., "Speech enhancement based on the subspace method", IEEE *Trans. Speech Audio Proc.*, vol.8, 2000, pp.497-507.

[3] Ephraim Y, "Statistical model-based speech enhancement systems", *Proc. IEEE*, vol.80, 1992, pp.1526-1558.

[4] Ephraim Y, Van Trees H L., "A signal subspace approach for speech enhancement", *IEEE Trans. Speech Audio Proc.*, vol.3, 1995, pp.251-266.

[5] Klemma V C, Laub AJ. "The singular value decomposition: its computation and some applications", *IEEE Trans. Automat. Control*, vol. 25, 1980, pp.164-176.

[6] Marple S L., "Digital spectral analysis with applications", Prentice Hall, Englewood Cliffs, NJ, 1987.

[7] Deller J R, Proakis J G, Hansen J H L., "Discrete-time Processing of Speech Signals", Maxwell McMillan, New York, 1993.

[8] Manfredi C., "Adaptive noise energy estimation in pathological speech signals", *IEEE Trans. Biomed. Eng.*, vol.47, 2000, pp.1538-1542.

[9] Manfredi C., D'Aniello M., Bruscaglioni P., "A simple subspace approach for speech denoising", *Logopedics Phoniatrics Vocology*, vol.26, p.179-192, 2001.

[10] Manfredi C., Landini L., Faita F., Gemignani V. SVD-based portable device for real-time hoarse voice denoising. Proc. Int. Conf. Digital Signal Processing, Santorini, GR, 2002, pp. 857-860.

[11] Hirano M., "Psycho-acoustic evaluation of voice", In: Hirano M. Clinical examination of voice, Springer-Verlag, New York, 1981.

[12] Golub G.H., Van Loan C.F., "Matrix Computations", 2nd Ed., Johns Hopkins University Press, 1989.

[13] Forsythe G.E., Malcolm M.A., Moler C.B., "Computer methods for mathematical computations", Prentice-Hall, 1977.

[14] Wilkinson J.H., Reinsch C., "Linear algebra", vol.II of handbook for automatic computation, Springler-Verlag, 1971.

[15] Stoer J., Bulirsch R., "Introduction to numerical analysis", Springer-Verlag, 1980.

[16] Press W. H., Flannery B. P., Teukolsky S. A., Vetterling W. T., "Numerical recipes in C – The art of scientific", Cambridge University Press, 1988.

[17] Dejonckere P H, Remacle M, Fresnel-Elbaz F, Woisard V, Crevier-Buchman L, Millet B, "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements", *Rev. Laryngol. Otol. Rhinol.*, vol.117, n.3, 1996, pp.219-224.