OPTIMISED COMB FILTER AND SVD FOR DYSPHONIC VOICE ENHANCEMENT

Claudia Manfredi, Cloe Marino

Department of Electronics and Telecommunications, Università degli Studi di Firenze
Via S.Marta 3, 50139 Firenze, Italy

manfredi@det.unifi.it

Abstract: Vocal fold paralysis, polyps, cordectomisation or other dysfunction, may alter regular speech production and cause more efforts to be used in speaking than for healthy people. In this paper, we deal with the problem of enhancing voice quality for people suffering from dysphonia, which is mainly due to air flow turbulence in the vocal tract, coming from irregular vocal folds vibration. We present a new approach for reducing voice hoarseness in pathological voices, often referred to as noise. Thanks to its robustness against noise, low-order singular value decomposition (SVD) of suitable data matrices is used for voice enhancement. An optimised adaptive comb filter (OACF) is applied first, to reduce noise between harmonics. Objective voice quality measures are proposed, to test results on real pathological data.

Keywords: Voice enhancement, SVD, Comb filter, Noise, pitch, spectrogram.

Introduction

The need for enhancing speech signals arises in many situations, especially in speech communication settings, in which the speech either originates from some noisy source or is affected by the noise at the receiving end [1], [3], [4]. At present, few results are available in the biomedical field, aiming at reducing voice hoarseness. However, this problem is of great concern, for rehabilitation and from the assistive technology point of view. Commonly, surgical and/or pharmacological treatments allow restoring voice quality, with patient's recovering to an acceptable or even excellent level. However, sometimes patients can only partly recover, with heavy implications on their quality of life.

The aim of the proposed method is to improve time and spectral characteristics of degraded voice signals. The method performs optimised adaptive comb filtering (OACF) on data windows of varying length, obtained with a new robust adaptive pitch estimation technique. OACF successfully reduces noise as evaluated by an adaptive implementation of the Normalised Noise Energy technique (ANNE). This step is followed by Singular value Decomposition (SVD) of matrices whose entries come from sampled speech data frames, properly organised. The noise component is removed from the signal and the filtered signal is reconstructed along the directions spanned by the eigenvectors associated with the signal subspace eigenvalues only, thus giving enhanced voice quality. New quality indexes have been defined, and used as objective measures for assessing enhancement of voice.

Real data (sustained vowels, words and sentences) coming from dysphonic subjects were filtered with the proposed approach, with enhanced results in almost all cases.

Method

Optimised Adaptive Comb Filter

An optimised adaptive comb filter (OACF) is applied first. The essence of comb filtering is to build a filter that passes the harmonics of the speech signal y(n), while rejecting noise frequency components between the harmonics [1], according to:

$$\hat{s}(n) = \sum_{i=-K}^{K} a(i)y(n-T_{0i}) \qquad (1)$$

Where T_{0i} is adaptively estimated according to the two steps described below, in sect.2.1.1 - 2.1.2. The filter that has been used in this paper has a Hamming window shape, which is obtained from the following equation (with K=3):

$$a(i) = \frac{0.54 + 0.46\cos(2\pi i/2K+1)}{\sum_{i=-K}^{K} 0.54 + 0.46\cos(2\pi i/2K+1)} \qquad (2)$$

Ideally, spacing between each "tooth" in the comb filter should correspond to F_0 in Hz, which, however, is often highly unstable in pathological voices. Hence, the comb filter must adapt to fast pitch variations. The filter is optimised here, as it relies on a robust adaptive pitch estimator allowing a varying window length for the filter, linked to varying pitch. To this aim, a two-step

ISSN: 1727-1983 © 2005 IFMBE

pitch detection algorithm is proposed. The choice of the techniques adopted in each step results from a comparative analysis of pitch extraction methods [2].

First Pitch Estimate

Simple Inverse Filter Tracking (SIFT) is applied first, on signal time windows of short but fixed length $M=3F_s/F_{min}$, where F_s is the signal sampling frequency, and F_{min} is the minimum allowed F_0 value for the signal under consideration (here: F_{min}=50Hz, corresponding to very low male pitch). Short time window is required, due to high non-stationarity of the signals under study. To create an IF, a low-order Linear Prediction (LP) is usually selected (order $p \approx 4$), since no more than two formants are expected in the low-passed signal frame [3]. For highly degraded signals as those under study, an adaptive choice for the filter order is proposed here, based on SVD of matrices whose entries come from sampled speech data frames [4], [5]. SVD requires selecting the "size" p of the signal subspace, i.e. the minimum number of eigenvectors spanning the clean data. To this aim, a variable threshold is defined, based on the Dynamic Mean Evaluation (DME) criterion, which relies on the geometric distance among "large" and "small" singular values [2]. The DME is applied to the decreasing sequence of singular values $\sigma^2{}_i$. Typically, with DME, $2 \le p \le 6$ during the utterance, due to changing signal characteristics: the larger the estimated p, the more varying the signal. From this step, a first raw F_0 tracking is obtained, along with its range of variation $[F_l, F_h]$.

Second Pitch Estimate

The second step gives a more accurate F_0 estimation and allows defining the optimum varying pitch period for OACF. F_0 is now adaptively estimated in the frequency range $[F_l, F_h]$, obtained in step 2.1.1. Estimation is performed on short time windows of varying length, corresponding to three pitch periods, inversely proportional to previously estimated local F_0, with 50% overlapping. The signal is band-pass filtered (50Hz-400Hz) with a proper Continuous Wavelet Transform (Mexican hat) and its periodicity is extracted by means of the Average Magnitude Difference Function (AMDF) approach, as non-stationarity and amplitude modulation of the signals under study often cause misestimation of the true signal periodicity with autocorrelation [2]. The procedure gives a sequence of F_{0i} pitch values, and the corresponding starting points in the time-domain T_{0i}, used for OACF (eqns. (1), (2)).

Singular Value Decomposition

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 [4],[5]. It performs the factorisation: $A=U\Sigma V^T$, T denoting transpose, for a matrix A, generally non-square. Matrix Σ is block-diagonal, with the (1,1) block given by: $\Sigma_p=\text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p)$, σ_i being the i-th singular value of A. The singular values σ_i display the distance of matrix A from low-rank matrices and together with the singular vectors U and V, they can be used to construct optimal low-rank approximants, A_p, where p is the size of the low-rank approximation. This considerably improves the quality of voice, removing a major source of sensitivity to noise. 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 data frames, whose length is adaptively obtained according to varying pitch period, as described in sect.2.1.

The proposed SVD-filtering method is based on the following steps [6]:

- Compute the SVD of $A = U\Sigma V^T = \sum_{k=1}^{r} \sigma_k u_k v_k^T$, $A \in \Re^{2(M-R)xR}$, $r=\min(R,2(M-R)) \ge p$. u_k and v_k are respectively the left and right singular vectors associated with the eigenvalue σ_k. R is choseh in the range: $F_s \le R \le M/2$, F_s = sampling frequency (kHz) and M = data frame length (number of points).
- Retain the p dominant singular values and the corresponding singular vectors, i.e.: $\hat{A}_p = \sum_{k=1}^{p} \sigma_k u_k v_k^T$. $\hat{A}_p$ is the p-rank approximation of A and 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 M-points speech frame is analysed. Filtered frames are put back together sequentially, appending the new frame to previously filtered frames.

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. A normalization step of the signal amplitude dynamics allows restoring the output level, lowered than the input one, due to scale factors in the filtering algorithm. Moreover, "click-noise" at the output, due to the filtering chain, was reduced with a linear interpolation across n=5 samples of the filtered signal, centred on the last sample of each frame. Despite its simplicity, the SVD approach was found effective in increasing voice quality [6], [7].

Noise estimation

An adaptive noise estimation technique is implemented, that allows tracking varying noise level during phonation. This in fact could be of help for the physician, in order to evaluate the effort made by the patient during the vocal emission. The ANNE (Adaptive Normalised Noise Energy), relies on the NNE comb filtering approach [8], optimised in order to deal with varying signal characteristics, as it is based on the two-step pitch detection algorithm described in sect. 2.1.1-2.1.2. ANNE is defined as:

$$\mathrm{ANNE}(k) = 10\log\left[\frac{\sum\limits_{m=N_L}^{N_H}\left|\tilde{W}_m(k)\right|^2}{\sum\limits_{m=N_L}^{N_H}\left|X_m(k)\right|^2}\right], k = N_L,\ldots,N_H \qquad (3)$$

with: $N_L=\lceil Nf_LT\rceil$, $N_H=\lceil Nf_HT\rceil$, 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, $|\tilde{w}_m(k)|^2$ = estimate of the unknown noise energy $|W_m(k)|^2$, $|X_m(k)|^2$ = signal energy, T = 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 . Hence, large negative ANNE values correspond to good voice quality, while values close to zero reflect the presence of noise. Due to their close relationship, the OACF was shown to perform a stronger noise reduction with respect to SVD, as far as ANNE is concerned, and is thus applied as pre-filtering step.

Objective Quality Indexes

Three objective indexes are defined¸ closely related to the signal characteristics. A frequency threshold value f_{th}=4kHz is defined, to separate the "harmonics" range from the "noise" one. It is based on the usual range for voiced sounds (first four formants) in adults [8], [9], as well as on experimental results obtained from threshold tuning in a dataset of voiced and unvoiced sounds. The subscript "non-filt" refers to the original signal, while "filt" refers to the denoised signal:

$$\mathrm{PSD}_{low} = 10\log_{10}\frac{\mathrm{PSD}_{non-filt}(f \le f_{th})}{\mathrm{PSD}_{filt}(f \le f_{th})} \qquad (4)$$

measures the ratio of the PSDs evaluated on the "harmonic range";

$$\mathrm{PSD}_{high} = 10\log_{10}\frac{\mathrm{PSD}_{non-filt}(f \ge f_{th})}{\mathrm{PSD}_{filt}(f \ge f_{th})} \qquad (5)$$

is the ratio of the PSDs, evaluated on the "noise range". A good denoising procedure should give PSD_{low} values near to zero (no loss of harmonic power), but high PSD_{high} values (loss of power due to noise). Finally, a measure of the denoising effectiveness (quality enhancement ratio, QER) is defined as:

QER is thus the ratio between the signal energy and that of the measured noise. QER>0 corresponds to good denoising.

$$\mathrm{QER} = 10\log_{10}\frac{\sum\limits_{n=1}^{M} y^2(n)}{\sum\limits_{n=1}^{M}(y(n) - y_{filt}(n))^2} \qquad (6)$$

Results

A set of about 20 voice signals (word /aiuole/) coming from adult male patients were analysed with the proposed approach. All patients underwent surgical removal of T1A glottic cancer, by means of laser or lancet technique. Perceptual evaluation with GRBAS scale showed good recovering, however, residual hoarseness was found in most of them. By applying OACF followed by SVD, voice quality results enhanced in most cases. The following figures are relative to one case (lancet). Figs.1-3 show the signal (upper plot), the two-step F_0 estimation (middle plot, black crosses for the first and the second estimate, respectively) along with its mean value and standard deviation (Std) the adaptive varying window length (middle plot, black stars), and the ANNE estimate (lowest plot, black crosses) with its mean value across the whole voiced emission.

Fig.1 is relative to the signal before denoising. F_0 is highly oscillating, with huge Std, as shown by unstable black crosses. Also, ANNE has almost low negative values, with a mean value of about -11.8dB.

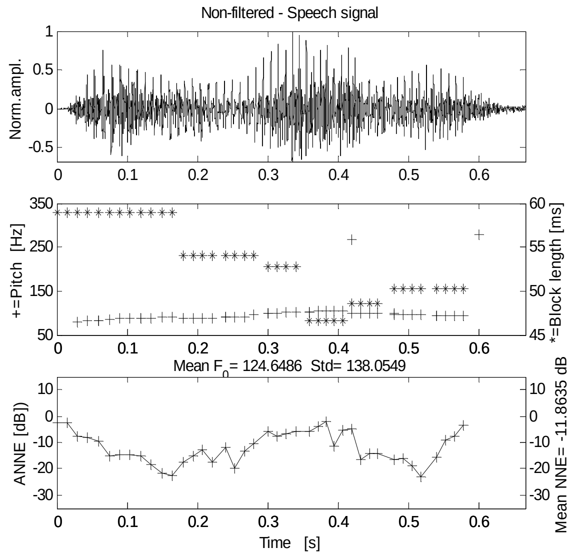

Figure 1: Non-filtered signal: two-step F_0 estimation, adaptive window length, ANNE

Figs.2 and 3 show the enhancement in F_0 stability (92.7Hz with low Std), and noise reduction after OACF and OACF+SVD: notice that after OACF the mean NNE value has decreased to about -18dB, with a small increase (-16.8dB) after OACF+SVD.

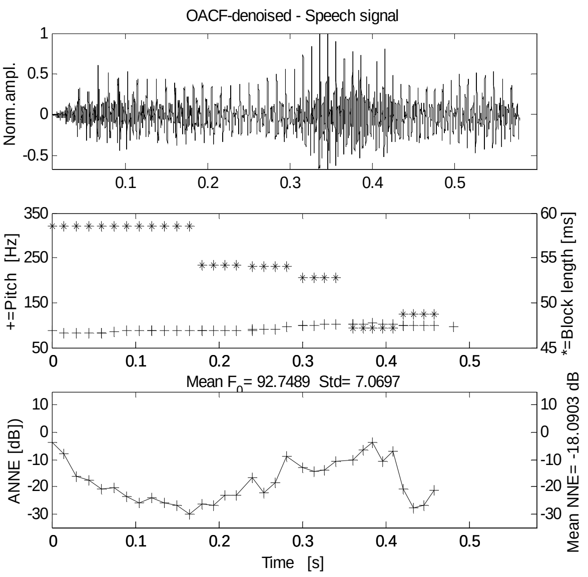

Figure 2: OACF-filtered signal: two-step F_0 estimation, adaptive window length, ANNE

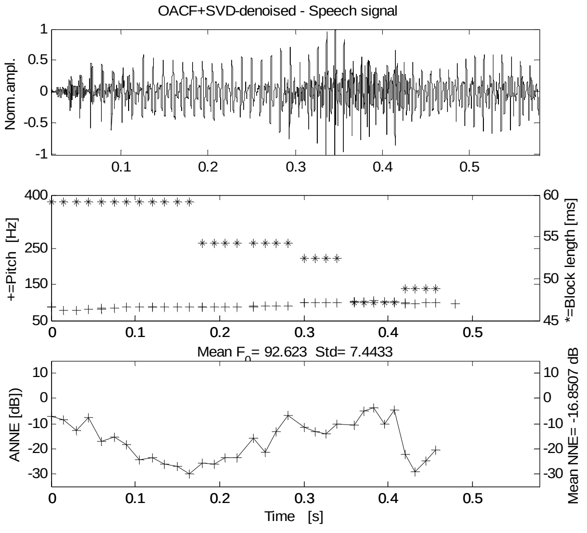

Figure 3: OACF+SVD-filtered signal: two-step F_0 estimation, adaptive window length, ANNE

Figs. 4-5 show the PSD plot and quality indexes (4)-(6) as obtained after OACF and OACF+SVD respectively. It is evident that OACF cannot remove enough noise energy in the spectrum, while after SVD a noticeable reduction is obtained. Specifically, for OACF, PSD_{low}=-0.22, PSD_{high}=2.19, QER=-2.8 indicate slight noise removal in the high frequency region, and even a small increase of noise energy in the spectrum. Strong denoising is obtained after SVD, with PSD_{low}=-1.84, PSD_{high}=17.2, QER=4.3. Notice that with SVD only, worse results are obtained.

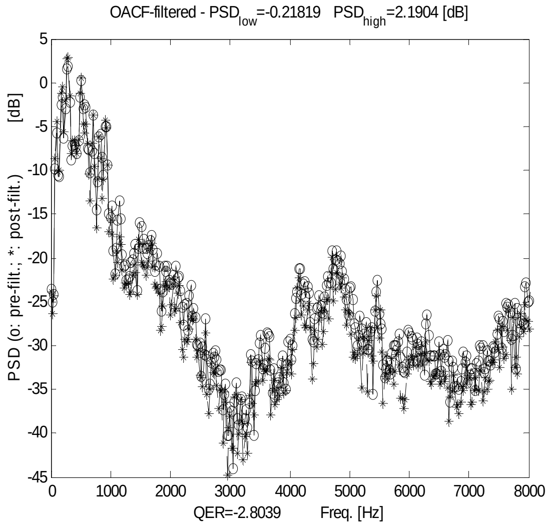

Figure 4: Non-filtered and OACF-filtered PSD plot along with new quality indexes

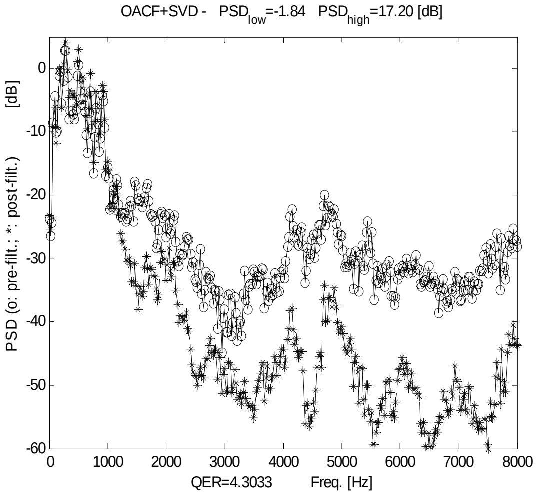

Figure 5: Non-filtered and OACF+SVD-filtered PSD plot along with new quality indexes

Finally, spectrograms in figs.7-8 clearly show noise reduction and harmonics enhancement after OACF, and especially OACF+SVD, as compared to the non-filtered signal spectrogram in fig.6 (the darker the grey level, the lower the spectral energy). Notice also that formants are preserved after denoising.

Summing up, first results show that OACF is well suited for noise removal between harmonics, while SVD is effective in filtering out high frequency noise. Combining the two approaches has given enhanced results in most cases.

Final remarks

A denoising procedure is proposed, based on an optimised ACF and low-order SVD decomposition of voice data. The procedure was found effective in increasing the quality of voice, while preserving the harmonic structure of the original signal. An automatic tool is provided, allowing robust pitch and noise estimation and strong noise reduction. This tool could be of help both for clinicians, in order to follow patient's rehabilitation, after surgery or drug treatment, and for dysphonic subjects, for testing and enhancing their fluent speech quality by means of a simple and cheap mobile device.A first prototype was implemented on a DSP board, by means of properly optimised C and Assembler code. The global procedure will be optimised, towards the fulfilment of a mobile device for real-time speech analysis and denoising.

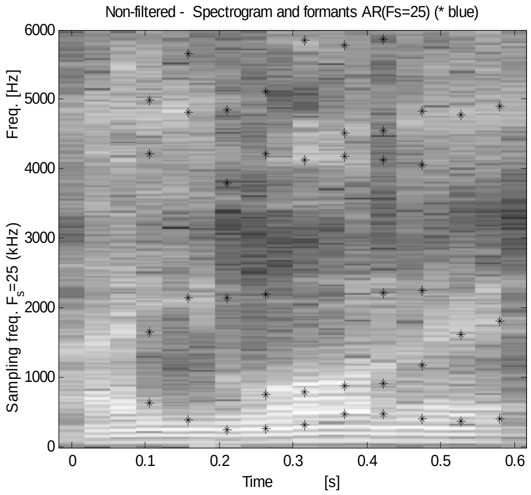

Figure 6: Signal spectrogram and formant tracking before denoising

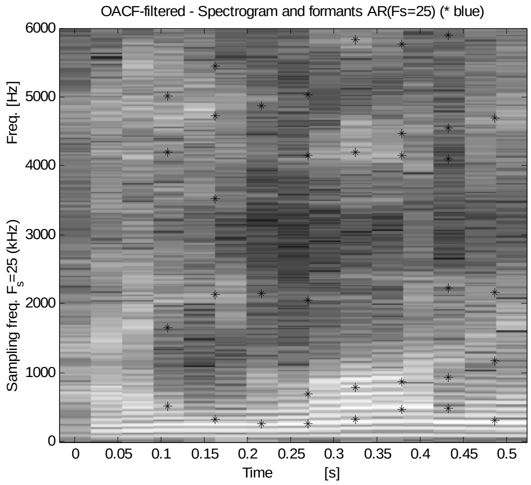

Figure 7: Signal spectrogram after OACF denoising

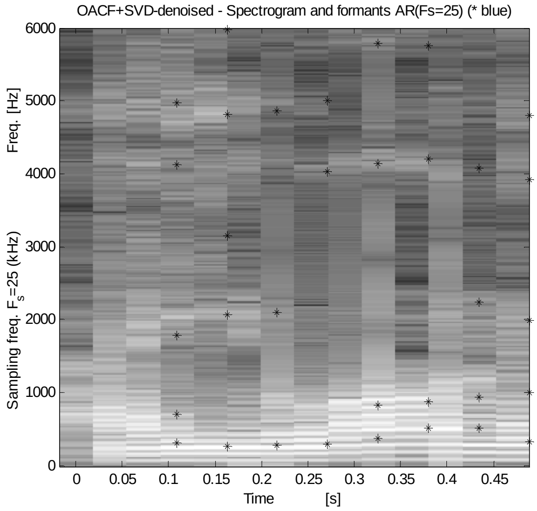

Figure 8: Signal spectrogram after OACF+SVD denoising

References

[1] LIM J.S., OPPENHEIM A.V., BRAIDA L.D., 'Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition', *IEEE Trans. Acoust.,Speech,Signal Proc.*, n.4, p.354-358, 1978.

[2] MANFREDI C., D'ANIELLO M., BRUSCAGLIONI P., ISMAELLI A., 'A Comparative Analysis of Fundamental Frequency Estimation Methods with Application to Pathological Voices', *Med. Eng. Phys.*, vol.22, n.2, pp.135-147, 2000.

[3] DELLER J R, PROAKIS J G, HANSEN J H L. *Discrete-time Processing of Speech Signals.* New York: Maxwell McMillan, 1993.

[4] EPHRAIM Y, VAN TREES H L. 'A signal subspace approach for speech enhancement'. *IEEE Trans.Speech Audio Proc.,*1995; n.3, p.251-266.

[5] RAO B D., ARUN K S. 'Model based processing of signals: a state space approach'. *Proc. IEEE* n.80, p.283-309, 1992.

[6] MANFREDI C., D'ANIELLO M., BRUSCAGLIONI P., 'A simple subspace approach for speech denoising', *Log. Phon. Vocol.*, vol.26, p.179-192, 2001.

[7] MANFREDI C., LANDINI L., FAITA F., SCHINDLER A., 'Voice quality enhancement with SVD', *2nd EMBEC Conf.*, vol.1, p.470-471,Vienna, Austria, 4-10 December, 2002.

[8] KASUYA H., OGAWA S., MASHIMA S., EBIHARA S., 'Normalised Noise Energy as an Acoustic Measure to Evaluate Pathologic Voice', *J. Acoust. Soc. Am.*, vol. 80, n.5, p.1329-1334, 1986.

[9] MANFREDI C., PERETTI G., 'A new insight into post-surgical objective voice quality evaluation. Application to thyroplastic medialisation', *IEEE Trans. Biom.Eng.,* 2005 (to appear).