ISCA Archive
http://www.isca-speech.org/archive

OPTIMISED GSVD FOR DYSPHONIC VOICE QUALITY ENHANCEMENT

Claudia Manfredi, Cloe Marino, Fabrizio Dori, Ernesto Iadanza

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

Abstract: **This paper concerns the problem of enhancing voice quality for people suffering from dysphonia, caused by airflow turbulence in the vocal tract, for irregular vocal folds vibration.**
A generalized subspace approach is proposed for enhancement of speech corrupted by additive noise, regardless of whether it is white or not. The clean signal is estimated by nulling the signal components in the noise subspace and retaining the components in the signal subspace. Two approaches are compared, taking into account both signal and noise, or signal only, eigenvalues. An optimised adaptive comb filter is applied first, to reduce noise between harmonics. Objective voice quality measures demonstrate improvements in voice quality when tested with sustained vowels or words corrupted with "hoarseness noise". The intention is to provide users (disabled people, as well as clinicians) with a device allowing intelligible and effortless speech for dysphonics, and useful information concerning possible functional recovering. This will be of use to people in social situations where they interact with non-familiar communication partners, such as at work, and in everyday life.

Keywords: hoarseness, voice denoising, GSVD, comb filtering, voice quality, pitch, noise, formants.

I. INTRODUCTION

Signal subspace methods are used frequently for denoising in speech processing, mainly with speech communication [1], [2]. Until now, few results are available concerning their application for voice quality enhancement in the biomedical field [3]. In this paper, the objective of noise reduction is to improve noisy signals due to irregular vocal folds vibration. 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 idea behind subspace methods is to project the noisy signal onto two subspaces: the signal subspace (since the signal dominates this subspace), and the noise subspace. The noise subspace contains signals from the noise process only, hence an estimate of the clean signal can be made by removing or nulling the components of the signal in the noise subspace and retaining only the components of the signal in the signal subspace. The decomposition of the space into two subspaces can be done using either the singular value decomposition (SVD) [4], [5] or the Quotient SVD (QSVD) or GSVD [1],[6],[11]. Though computationally expensive, GSVD was found robust and effective in reducing noise due to turbulences in the vocal tract, which is typically coloured. GSVD is implemented here with two choices for separating the signal and the noise subspaces, to compare performance. Specifically, the first choice is based on classical GSVD, where both the signal and the noise subspace eigenvalues are used for filtering [6]. The second one corresponds to retaining the signal subspace eigenvalues only [1].
An adaptive comb filter is applied first, as it was shown to significantly reduce noise between the harmonics in the spectrum. The comb filter is optimised, in the sense that it is applied on windows whose length varies according to varying pitch.
Real data coming from dysphonic subjects are successfully denoised with the proposed approaches.

II. MATERIALS AND METHODS

Firstly, optimised adaptive comb filtering is performed on data windows of varying length, obtained with a new two-step robust adaptive pitch estimation technique [7]. The essence of comb filtering is to build a filter that passes the harmonics of the noisy speech signal y, while rejecting noise frequency components between the harmonics [8],[9]. Ideally, spacing between each "tooth" in the comb filter should correspond to F_0 ($1/T_0$) in Hz, which is often highly unstable in pathological voices. The proposed comb filter, based on an adaptive two-step pitch estimator, is capable to adapt to fast pitch variations and successfully reduces noise as evaluated by an adaptive implementation of the Normalised Noise Energy technique (ANNE) [7], thus being suited as a pre-filtering step. The filter that has been used in this paper has a Hamming window shape, which is obtained from the following equation (with K=3):

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

Models and analysis of vocal emissions for biomedical applications. 4th international workshop. October 29-31, 2005 – Firenze, Italy.
Edited by C. Manfredi. ISBN 88-8453-320-1 (online) © 2005 Firenze University Press

This step is followed by Generalised Singular Value Decomposition (GSVD) of signal and noise matrices, whose entries are suitably organised, as shown in eq. (4). GSVD-based voice denoising aims at diminishing the uncorrelated and added noise from the voice signal, weather it is white or not. The noisy signal y at time instant t, y_t ,can be expressed as:

$$y_t = d_t + n_t \quad (2)$$

Where d=clean signal, n=(coloured) noise. The goal is to estimate d from y. The noisy signal is segmented into frames y_i, i=1, 2,…, of varying length M_i, obtained according to the previously cited robust adaptive pitch estimation procedure. The GSVD amounts to finding a non-singular matrix X and two orthogonal matrices U, V of compatible dimensions, which simultaneously transform both the Hankel noisy speech matrix H_y and the noise matrix H_n into nonnegative diagonal form matrices C and S such as:

$U^T H_y X = C = \mathrm{diag}(c_1,\ldots,c_k)$, $c_1 \geq c_2 \geq \ldots \geq c_K$

$V^T H_n X = S = \mathrm{diag}(s_1,\ldots,s_k)$, $s_K \geq s_{K-1} \geq \ldots \geq s_1$

$C^T C + S^T S = I_K$ (3)

Where L+K=M+1, K<L. The H_y matrix has the form:

$$H_y = \begin{bmatrix} y_0 & y_1 & \cdots & y_{K-1} \\ y_1 & y_2 & \cdots & y_K \\ \vdots & \vdots & \ldots & \vdots \\ y_{L-1} & y_L & \cdots & y_{M-1} \end{bmatrix} \quad (4)$$

Similarly for H_n.

The values $c_1/s_1 \geq c_2/s_2 \geq \ldots, \geq c_K/s_K$ are referred to as the generalised singular values of H_y and H_n. Notice that one can choose to work with Toepliz matrices instead of Hankel matrices. There are no fundamental differences between the two approaches.

It was shown [1], [2], [6], [11] that the filtered signal can be obtained either from the matrix:

$$H_y^p = U \begin{bmatrix} C_p S_p^{-1} & 0 \\ 0 & 0 \end{bmatrix} X^{-1} \quad (5)$$

or from the matrix:

$$H_y^p = U \begin{bmatrix} C_p & 0 \\ 0 & 0 \end{bmatrix} X^{-1} \quad (6)$$

where U and X are as in eq. (3) and $C_p=\mathrm{diag}(c_1,\ldots, c_p)$, $S_p= \mathrm{diag}(s_1,\ldots, s_p)$, are sub-matrices of C and S respectively and p is the signal subspace dimension. Eq. (5) corresponds to classical GSVD, where both the signal and the noise subspace eigenvalues are used for filtering, and will be referred to as GSVD in what follows. Eq. (6) corresponds to retaining the signal subspace eigenvalues only, and will be referred to as OSV (Only Signal Values).

Two problems were encountered with GSVD, i.e. the choice of the noise covariance matrix and that of the signal subspace dimension p. Commonly, in speech communication settings, the noise covariance matrix is computed using noise samples collected during speech-absent frames. To deal with the problem under study, different choices were tested. Among them, one takes into account the signal noisy component as obtained from a preliminary SVD decomposition of the signal under study: the noise subspace is reconstructed and used to fill matrix H_n. While giving almost good results, this choice was disregarded, due to both the larger computational load and to better results obtained with the following approach: on each signal frame of varying length, an AutoRegressive (AR) model is identified, and the model residuals are evaluated. The residual variance is then used to construct the diagonal matrix S of eq. (3).

The second problem is the optimal choice of the number p of retained singular values for denoised signal reconstruction. Classical order selection criteria were applied to GSVD, such as AIC, MDL [9], as well as a new criterion named DME [10], but best results were obtained with p=2. It will be named as GSVD_{fix} in what follows. As for OSV, p was chosen such as [1]:

$$c_p > s_p \text{ and } c_{p+1} < s_{p+1} \quad (7)$$

This was in fact the choice that gave the best results.

Finally, three objective indexes are defined, closely related to the signal characteristics. A frequency threshold value f_{th}=4kHz is defined, based on the usual range for voiced sounds (first four formants) in adults, 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:

$$\mathrm{PSD}_{low} = 10\log_{10} \frac{\mathrm{PSD}_{non-filt}(f \leq f_{th})}{\mathrm{PSD}_{filt}(f \leq f_{th})} \quad (8)$$

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

$$\mathrm{PSD}_{high} = 10\log_{10} \frac{\mathrm{PSD}_{non-filt}(f \geq f_{th})}{\mathrm{PSD}_{filt}(f \geq f_{th})} \quad (9)$$

is the ratio of the PSDs, evaluated on the "noise range". A

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

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 removed noise. QER>0 corresponds to good denoising [10].

III. 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 glottis cancer, by means of laser or lancet technique. Perceptual evaluation with GIRBAS scale showed good recovering, however, residual hoarseness was found in

most of them. By applying the proposed adaptive comb filter, followed by GSVD or OSV, voice quality results enhanced in most cases. The following figures are relative to one case (lancet operated). Each plot shows F_0, noise and formants tracking, as obtained by means of the cited robust, adaptive, high-resolution tool, along with F_0 and noise mean values.

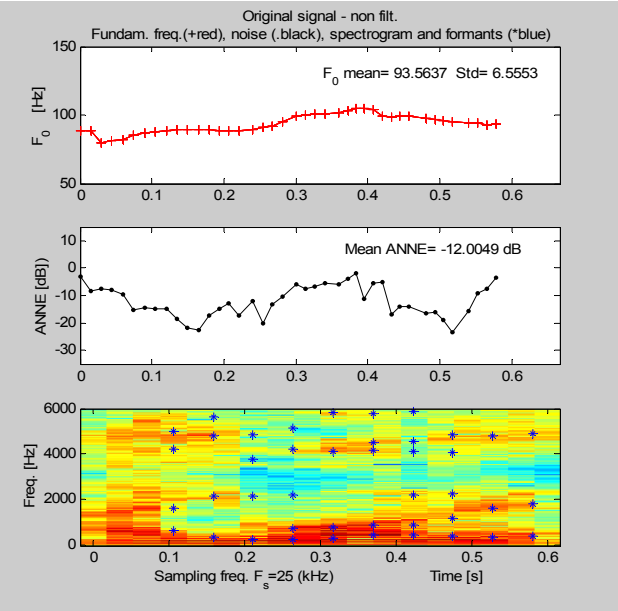

Figure 1 – Non-filtered signal: F_0, noise and formant tracking (superimposed on the spectrogram). High noise level is found, also in the high-frequency region.

Specifically, fig.1 is relative to the non-filtered signal: F_0 is almost stable, but the harmonics noise level is high (around -12 dB). The spectrogram shows strong noise also in the high-frequency region.
Fig.2 concerns comb-filtered signal. It shows still stable F_0, but harmonics noise is now lowered (from -12dB to about -18 dB). In the spectrogram, lower noise energy is shown also in the high-frequency spectral region.
Fig.3 refers to the signal filtered with comb and $GSVD_{fix}$. Harmonics noise is slightly raised (from -18 dB to about -14 dB), but the spectrogram evidences very low noise in the high frequency region.
Finally, fig.4 shows the results obtained for the signal filtered with comb and OSV with signal subspace as from eq.(7). Harmonics noise is lower than with GSVD (around -16.5 dB) and the spectrogram results comparable to the GSVD one. In all the figures (1)-(4) formant tracking is also reported, showing that the harmonics structure of the original signal is preserved with filtering.
The last fig.5 compares the values of PSD_{low}, PSD_{high} and QER for the applied denoising techniques, specifically comb, comb+$GSVD_{fix}$, comb+OSV, relative to the non-filtered signal. Best results are obtained with comb+$GSVD_{fix}$. As shown in the figure, comb alone performs only a slight enhancement, while comb+$GSVD_{fix}$ gives the best results with respect to other methods, with PSD_{low} ≅0dB, PSD_{high}>>0, and QER<0. Notice that previous results obtained with SVD_{fix} gave: Mean F_0=97.8Hz with std=28.6Hz, mean ANNE=-11.1dB PSD_{low}=-2.1 dB, PSDhigh=16.5 dB, QER=3.1 dB [3]. With comb + SVD, we obtained: Mean F_0= 92.6 with std= 7.3, mean ANNE=-16.5dB, PSD_{low}=-1.8dB, PSD_{high}=17.2dB, QER=4.3 dB.

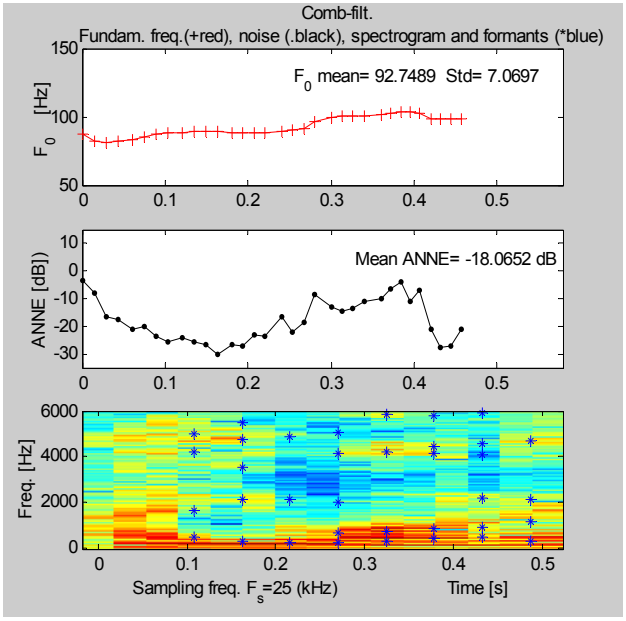

Figure 2 – Comb-filtered signal: F_0, noise and formant tracking (superimposed on the spectrogram). Harmonics noise is lowered (from -12dB to about -18 dB).

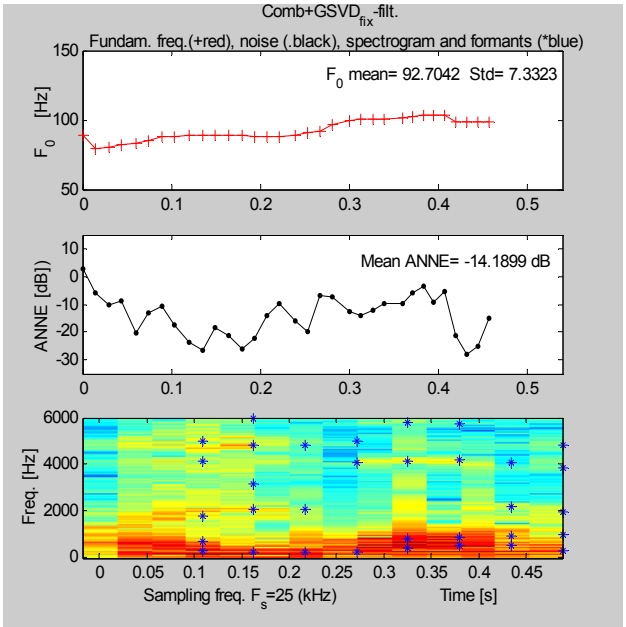

Figure 3 – Signal filtered with comb and $GSVD_{fix}$. F_0, noise and formant tracking (superimposed on the spectrogram). The spectrogram evidences lowered noise in the high frequency region.

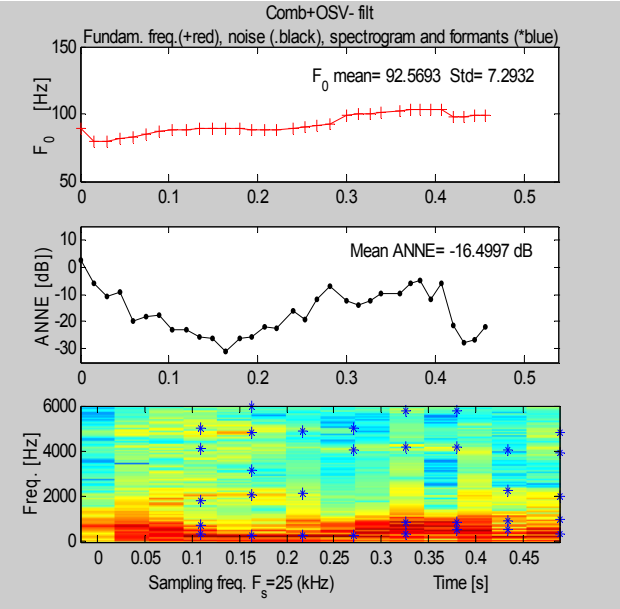

Figure 4 - Signal filtered with comb and OSV with signal subspace as from eq.(7): F_0, noise and formant tracking (superimposed on the spectrogram). Spectrogram comparable to fig.3.

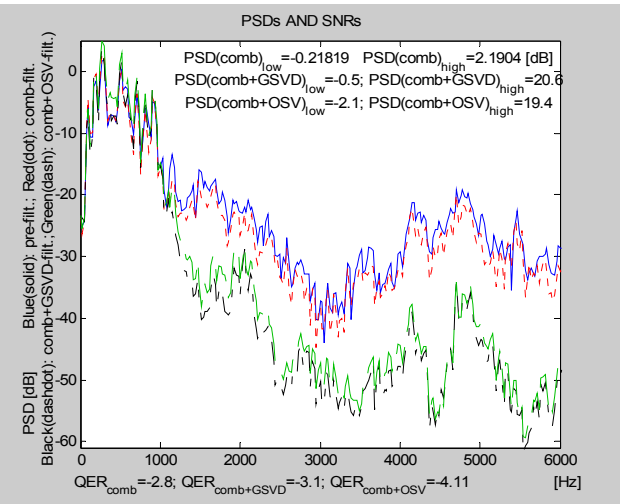

Figure 5 – Comparison among PSD and QER values obtained from eqs. (8)-(10) for comb, comb+$GSVD_{fix}$, comb+OSV, related to the non-filtered signal. Best results are obtained with comb+$GSVD_{fix}$.

This means that with SVD alone F_0 becomes more unstable and harmonics noise is increased. By pre-filtering with adaptive comb, results become comparable to comb+$GSVD_{fix}$ and comb+OSV, although a little bit worse. Similar results were obtained over all the dysphonic voices data set.

IV. FINAL REMARKS

A hoarse voice denoising procedure is proposed, based on an optimised comb filtering and low-order GSVD decomposition of voice data matrices. An automatic tool is provided, for robust pitch, noise and formant tracking. The whole procedure was found effective in increasing the quality of voice, as measured by few but effective objective indexes, while preserving the harmonic structure of the original signal. A perceptual comparison of results with GIRBAS scale will be available in the next future.

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. As a drawback, GSVD has a significant computational load, and for time being it is only used as an off-line algorithm. Recursive updating of GSVD, instead of re-computing it on each data window, would be desirable for real-time voice signal processing and is a topic of current research.

REFERENCES

[1] Jensen S., Hansen P., Hansen S., Sorensen J., "Reduction of broad-band noise in speech by truncated QSVD", *IEEE Trans. on SAP*, vol.3, p.439-448, 1995.

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

[3] Manfredi C., D'Aniello M., Bruscaglioni P., "A simple subspace approach for speech denoising", *Log. Phon. Vocol.*, vol.26, p.179-192, 2001.

[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] Ju G., Lee L., "Speech enhancement based on Generalised Singular Value Decomposition approach", *Proc.ICSLP 2002*, p.1801-1804, 2002.

[7] Manfredi C. Adaptive noise energy estimation in pathological speech signals. *IEEE Trans. Biomed. Eng*. 2000; 47: 1538-1542.

[8] 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.

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

[10] 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).

[11] Hu Y., Loizou P.C., "A generalised subspace approach for enhancing speech corrupted by coloured noise", *IEEE Trans. Speech Audio Proc.*, vol.11, p.334-341, 2003.