Efficient GCI detection for efficient sparse linear prediction

Vahid Khanagha, Khalid Daoudi

Efficient GCI detection for efficient sparse linear prediction

Vahid Khanagha and Khalid Daoudi

INRIA Bordeaux Sud-Ouest (GEOSTAT team)
200 Avenue de la vielle tour, 33405 Talence, France
email: vahid.khanagha@inria.fr, khalid.daoudi@inria.fr
http://geostat.bordeaux.inria.fr/

Abstract. We propose a unified non-linear approach that offers an efficient closed-form solution for the problem of sparse linear prediction analysis. The approach is based on our previous work for minimization of the weighted l_2-norm of the prediction error. The weighting of the l_2-norm is done in a way that less emphasis is given to the prediction error around the Glottal Closure Instants (GCI) as they are expected to attain the largest values of error and hence, the resulting cost function approaches the ideal l_0-norm cost function for sparse residual recovery. As such, the method requires knowledge of the GCIs. In this paper we use our recently developed GCI detection algorithm which is particularly suitable for this problem as it does not rely on residuals themselves for detection of GCIs. We show that our GCI detection algorithm provides slightly better sparsity properties in comparison to a recent powerful GCI detection algorithm. Moreover, as the computational cost of our GCI detection algorithm is quite low, the computational cost of the overall solution is considerably lower.

1 Introduction

It is generally desirable for voiced speech sounds that the residuals obtained by their Linear Prediction (LP) analysis (the prediction error) form a sparse time series that is almost zero all the time, except for few instants where considerably larger values of error may occur (the outliers). Such sparse representation is beneficial for coding applications [9, 13] and also leads in a more accurate estimation of human vocal tract system [11, 16].

Classical approaches for LP analysis of speech rely on minimization of l_2-norm of prediction error. As the l_2-norm is numerically highly sensitive to outliers [18], such minimum variance solution favors non-sparse solutions with many small non-zero entries rather than the sparse solutions having the fewest possible non-zero entries. A more suitable cost function is the l_1-norm of prediction error as it puts less emphasis on outliers. The l_1-norm minimization of residuals is already shown to be advantageous for sparse residual recovery with the use of convex programming tools with special care to avoid an unstable solution [5, 10, 12].

We have previously shown in [16] that the minimization of a weighted version of l_2-norm results in a more sparse solution with less computational burden. In this approach, GCIs are considered as the points that have the potential of attaining largest norms of residuals and then a weighting function is constructed such that the prediction error is relaxed at those points. Consequently, the weighted l_2-norm objective function is minimized by the solution of normal equations of liner least squares problem. In [16] a powerful method called SEDREAMS [7] is employed for detection of GCIs and it is shown that the weighted l_2-norm minimization achieves better sparsity properties compared to l_1-norm minimization.

In continuation of our sparse weighted l_2-norm solution mentioned above, in this paper we employ our recently developed GCI detection algorithm [15, 17] for construction of the weighting function. This GCI detection algorithm relies on a novel multi-scale non-linear signal processing approach called the Microcanonical Multiscale Formalism (MMF) and extracts GCIs directly from the speech signal without being based on any model of speech production. This makes it more suitable to be used for sparse residual recovery compared to SEDREAMS. Indeed, although SEDREAMS is shown to provide the best of performances compared to several state-of-the-art algorithms [7], the fact that it relies on residuals themselves for detection of GCIs, controverts its use for sparse residual recovery (SEDREAMS uses the classical minimum variance solution for calculation of LP residuals). We show that our GCI detection algorithm provides slightly better sparsity properties compared to SEDREAMS when used inside the weighted-l_2-norm solution. Moreover we show that the MMF-based solution is more efficient than SEDREAMS and hence is a better choice for keeping down the overall computational burden.

The paper is organized as follows. In section 2 we present the weighted l_2-norm minimization approach. We then briefly introduce in section 3, the MMF based GCI detection algorithm that we use in this paper. In section 4, the experimental results are presented and finally in section 5, we draw our conclusion and perspectives.

2 The weighted l_2-norm solution

The ideal solution for sparse residual recovery in LP analysis is to directly minimize the number of non-zero elements of the residual vector, i.e. its cardinality or the so-called l_0-norm [4]. This minimization problem is however a N-P hard optimization problem [12]. The classical LP analysis technique on the other hand is very simple (minimization of l_2-norm of residuals), but can not provide the desired level of sparsity. This is due to the exaggerative effect of l_2-norm on larger values of error (the so-called outliers) [18]. Indeed, the minimizer puts its effort on lowering the value of these outliers, with the cost of more non-zero elements and hence the resulting residuals are not as sparse as desired. To reduce the exaggerative effect of l_2-norm on the outliers, one may replace it with l_1-norm which is less sensitive to larger values [4]. The l_1-norm minimization problem can be solved by by recasting the minimization problem into a linear program [3] and then using convex optimization tools [2]. Special care must be taken however, to avoid stability and computational issues for the case of speech signal [5, 10].

We have shown in [16] that a simple minimization of the weighted version of l_2-norm cost function may lead to better sparsity properties, while avoiding stability issues. Indeed, the use of l_2-norm cost function preserves the computational efficiency while the weighting function is used to cope with its exaggerative effect on outliers (by careful down-weighting of the cost function at those points).

Formally, the linear prediction of the speech signal $x(n)$ with a set of K prediction coefficients a_k can be written as:

$$x(n) = \sum_{k=1}^{K} a_k x(n-k) + r(n) \tag{1}$$

The weighted l_2-norm solution is based on solving the following optimization problem:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \quad \sum_{k=1}^{N} w(k)(r(k)^2) \tag{2}$$

where $w(\cdot)$ is the weighting function. Once $w(\cdot)$ is properly defined, the solution to Eq. (2) is straight-forward. Indeed, setting the derivative of the cost function to zero results in a set of normal equations that can be easily solved as explained in [16].

The weighting function is constructed in a way that it de-emphasizes the exaggerative effect of l_2-norm on outliers. This is achieved by down-weighting the l_2-norm at the time instants where an outlier is expected to occur. Indeed, it is argued in [16] that the outliers of LP analysis are expected to coincide with GCIs. Hence, the weighting function is expected to provide the lowest weights at the GCI points and to give equal weights to the remaining points. To put a smoothly decaying down-weighting around GCI points and to have a controllable region of tolerance around them, a natural choice is to use a Gaussian-shape weighting function. The final weighting function is thus defined as:

$$w(n) = 1 - \sum_{k=1}^{N_{gci}} g(n - T_k) \tag{3}$$

where $T_k, k = 1 \cdots N_{gci}$ denotes the detected GCI points and $g(\cdot)$ is a Gaussian function ($g(x) = \kappa e^{(\frac{x}{\sigma})^2}$). The parameter σ allows the control of the width of the region of tolerance and κ allows the control of the amount of down-weighting on GCI locations. Overall, in minimizing the resulting weighted l_2-norm cost function (Eq. (2)), the minimizer is free to pick the largest residual values for the outliers and it concentrates on minimizing the error on the remaining points and hence, the sparsity is granted. It is noteworthy that a recent work (independently) proposed a similar weighted linear prediction which downgrades the contribution of the glottal source and, thus, leads to more accurate formant estimates [1]. This model does not use however GCI estimation.

3 The MMF based GCI detection algorithm

We have introduced a robust GCI detection algorithm in [15, 17] that is based on a novel multi-scale non-linear signal processing approach called the MMF. The MMF is centered around precise estimation of local quantities called the Singularity Exponents (SE) at every point in the signal domain. When correctly defined and estimated, these exponents alone can provide valuable information about the local dynamics of complex signals. The singularity exponent $h(n)$ for any given signal $x(n)$, can be estimated by evaluation of the power-law scaling behavior of a multi-scale functional $\mathit{\Gamma}_r$ over a set of fine scales r:

$$\mathit{\Gamma}_r\left(x(n)\right) \propto r^{h(n)} + \mathrm{o}\left(r^{h(n)}\right) \qquad r \to 0 \tag{4}$$

where $\mathit{\Gamma}_r\left(\cdot\right)$ can be any multi-scale functional complying with this power-law like the gradient-based measure introduced in [21]. The details about the choice of $\mathit{\Gamma}_r\left(\cdot\right)$ and the consequent estimation of $h(n)$ are provided in [21, 15].

In MMF, a particular set of interest is the level set called the *Most Singular Manifold* (MSM) which comprises the points having the smallest SE values. It has been established that the critical transitions of a complex signal occurs at these points. This property has been successfully used in several signal processing applications [19, 20, 22]. In case of voiced speech sounds, we have shown in [15, 17] that indeed the MSM coincides with the instants where the most significant excitations of the vocal tract occur (the GCIs). Consequently, we used this property to develop a simple GCI detection algorithm which is much faster than SEDREAMS and we showed that it has almost the same performance as the SEDREAMS algorithm in case of the clean speech. In case of noisy speech our MMF-based approach shows considerably higher accuracy in comparison to the SEDREAMS in presence of 14 different types of noise taken from different real-world acoustic environments.

In this paper we aim at using this MMF-based GCI detection algorithm inside the weighted l_2-norm solution of section 2 for sparse residual recovery. Indeed, the MMF-based algorithm is particularly suited for this problem as it extracts the GCIs exclusively using the singularity exponents $h(t)$ and as opposed to SEDREAMS, it does not rely on residuals themselves for GCI detection. Moreover, as the computational complexity of our MMF-based GCI detector is much less than SEDREAMS and hence, the overall computational effort for solution of Eq. (2) will be reduced. As such, this MMF-based GCI detector can be seen as a complement to the weighted l_2-norm solution of section 2 for sparse residual recovery: combined together, they form a unified efficient and closed-form solution for sparse LP analysis of speech.

4 Experimental results

We perform an extensive set of experiments on 3000 utterances that are randomly selected from TIMIT database to compare the sparsity properties of the residuals obtained by four different LP approaches: the classical l_2-norm minimization, the l_1-norm minimization, weighted l_2-norm minimization using SEDREAMS for GCI detection and weighted l_2-norm minimization using our MMF-based GCI detector. Following [9], the utterances are downsampled to 8 KhZ and the prediction order of $K = 13$ is used. As for the l_1-norm minimization method, we use the publicly available l_1-magic toolbox [3] which uses the primal-dual interior points optimization [2]. For the weighted l_2-norm minimization, all the results presented in this section are obtained using the following set of parameters for $w(\cdot)$: $\kappa = 0.9$ and $\sigma = 50$. The choice of the parameters was obtained using a small development set (of few voiced frames) taken from the TIMIT database [8]. As for the SEDREAMS GCI detection algorithm [7] we use the implementation that is made available on-line by its author [6] (GLOAT toolbox).

Fig. 1 shows an example of the residuals obtained by all these different optimization strategies. It is clear that the weighted-l_2 and also l_1-norm criteria achieve higher level of sparsity compared to the classic l_2-norm criterion. Moreover, a closer look reveals that in this example the weighted-l_2-norm solution which uses the MMF based GCI detector shows better sparsity properties compared to the rest of optimization strategies.

To perform a more quantitative comparison, we use the kurtosis of residuals as a quantitative measure of the level of sparsity of the obtained residuals. The kurtosis is an appropriate measure of sparsity as it satisfies three of the most important properties that are intuitively expected from such a measure: scale invariance, rising tide and Robin Hood [14]. Kurtosis is a measure of peakedness of a distribution and higher values of kurtosis implies higher level of sparsity. Table 1 shows the kurtosis of the residuals obtained from the above-mentioned optimization strategies, averaged over the 3000 utterances from TIMIT database. It can be seen from table 1 that the highest value for the kurtosis is achieved by the weighted l_2-norm solution, when the MMF based approach is used for GCI detection.

Table 1. Quantitative comparison of sparsity level of different LP analysis strategies.

Method	kurtosis on the whole sentence	kurtosis on voiced parts
l_2-norm	49.46	95.67
l_1-norm	67.01	114.93
weighted-l_2-norm+SEDREAMS GCIs	62.78	114.69
weighted-l_2-norm+MMF GCIs	66.23	120.62

In terms of computational efficiency we compare the computational processing times for the weighted l_2-norm solutions using SEDREAMS and MMF for GCI detection, in terms of the average empirical Relative Computation Time

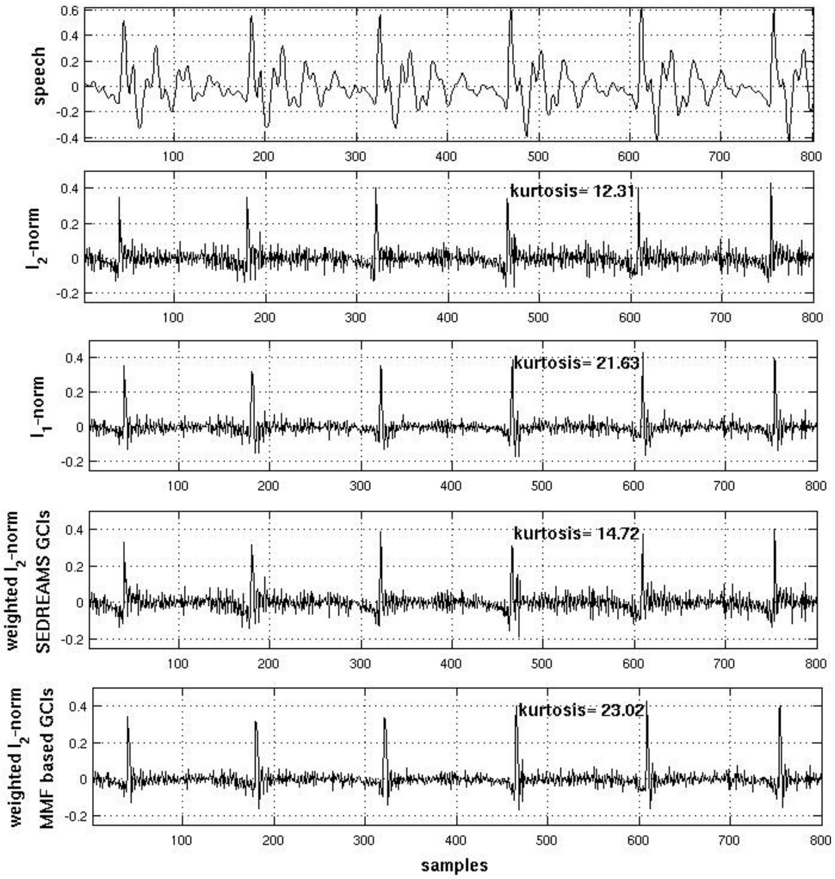

Fig. 1. The residuals of the LP analysis obtained from different optimization strategies for Fs=16 Khz, K = 26 and frame length of 25 msec.

that is defined as:

$$RCT(\%) = 100.\frac{CPU\ time\ (s)}{Sound\ duration\ (s)} \tag{5}$$

The results are reported in table 2 and it can be seen that the MMF based solution is indeed much faster than the one based on SEDREAMS. It must be noted that we have used the original implementation of SEDREAMS in the GLOAT toolbox [6] and not its fast implementation which should be about 4 times faster than the original one [7].

Table 2. Comparison of the average Relative Computation Time.

Method	RCT (%)
weighted-l_2-norm+SEDREAMS GCIs	46.97
weighted-l_2-norm+MMF GCIs	7.75

5 Conclusions

In continuation of our previous work for retrieval of sparse LP residuals using a weighted l_2-norm minimization [16], in this paper we introduced the use of our MMF based GCI detection algorithm for construction of the weighting function. Apart from having considerably lower computational cost, this GCI detection algorithm is particularly suited for sparse residual recovery as it does not rely on residuals themselves and extracts GCIs exclusively from geometric multi-scale measurements. As such, the MMF based GCI detection serves as a complement to our weighted-l_2-norm solution for sparse residual recovery. We showed that such a unified approach provide slightly better sparsity properties, while considerably reduces the overall computational burden of the algorithm. These results suggest further investigation of the potential of such a unified non-linear solution for speech analysis: performance in presence of noise, the accuracy of formant estimation, the quality of the synthesis when used inside parametric multipulse coders and etc. This indeed will be the subject of our future communications.

References

1. Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A., Story, B.: Improved formant frequency estimation from high-pitched vowels by downgrading the contribution of the glottal source with weighted linear prediction. In: INTERSPEECH (2012)
2. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
3. Candès, E.J., Romberg, J.: l1-magic : Recovery of sparse signals via convex programming, 2005

4. Candès, E.J., Wakin, M.B.: Enhancing sparsity by reweighted l1 minimization. Journal of Fourier Analysis and Applications 14, 877–905 (2008)
5. Denoel, E., Solvay, J.P.: Linear prediction of speech with a least absolute error criterion. IEEE Transactions on Acoustics, Speech and Signal Processing, 33, 1397–1403 (1985)
6. Drugman, T.: Gloat toolbox. [Online], http://tcts.fpms.ac.be/ drugman/
7. Drugman, T., Thomas, M., Gudnason, J., Naylor, P., Dutoit, T.: Detection of glottal closure instants from speech signals: A quantitative review. IEEE Transactions on Audio, Speech, and Language Processing 20(3), 994 –1006 (march 2012)
8. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V.: DARPA TIMIT acoustic-phonetic continuous speech corpus. Tech. rep., U.S. Dept. of Commerce, NIST, Gaithersburg, MD (1993)
9. Giacobello, D.: Sparsity in Linear Predictive Coding of Speech. Ph.D. thesis, Multimedia Information and Signal Processing, Department of Electronic Systems, Aalborg University (2010)
10. Giacobello, D., Christensen, M.G., Dahl, J., Jensen, S.H., Moonen, M.: Sparse linear predictors for speech processing. In: Proceedings of the INTERSPEECH (2009)
11. Giacobello, D., Christensen, M.G., Murth, M.N., Jensen, S.H., Marc Moonen, F.: Sparse linear prediction and its applications to speech processing. IEEE Transactions on Audio, Speech and Language Processing 20, 1644–1657 (2012)
12. Giacobello, D., Christensen, M.G., Murthi, M.N., Jensen, S.H., Moonen, M.: Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2010)
13. Giacobello, D., Christensen, M., Murthi, M., Jensen, S., Moonen, M.: Retrieving sparse patterns using a compressed sensing framework: Applications to speech coding based on sparse linear prediction. IEEE Signal Processing Letters 17 (2010)
14. Hurley, N., Rickard, S.: Comparing measures of sparsity. IEEE Transactions on Information Theory 55, 4723–4740 (2009)
15. Khanagha, V.: Novel Multiscale methods for non-linear speech analysis., [Online] http://geostat.bordeaux.inria.fr/index.php/vahid-khanagha.html. Ph.D. thesis, University of Bordeaux 1 (2013)
16. Khanagha, V., Daoudi, K.: An efficient solution to sparse linear prediction analysis of speech. EURASIP Journal on Audio, Speech, and Music Processing (2013)
17. Khanagha, V., Daoudi, K., Yahia, H.: A novel multiscale method for detection of glottal closure instants. submitted to IEEE Transactions on Audio, Speech, and Language Processing (2013)
18. Meng, D., Zhao, Q., Xu, Z.: Improved robustness of sparse pca by l1-norm maximization. Pattern Recognition Elsevier 45, 487–497 (2012)
19. Turiel, A., del Pozo, A.: Reconstructing images from their most singular fractal manifold. IEEE Transactions on Image Processing 11, 345–350 (2002)
20. Turiel, A., Parga, N.: The multi-fractal structure of contrast changes in natural images: from sharp edges to textures. Neural Computation 12, 763–793 (2000)
21. Turiel, A., Yahia, H., Pérez-Vicente., C.: Microcanonical multifractal formalism: a geometrical approach to multifractal systems. part 1: singularity analysis. Journal of Physics A: Mathematical and Theoretical 41, 015501 (2008)
22. Yahia, H., Sudre, J., Garçon, V., Pottier, C.: High-resolution ocean dynamics from microcanonical formulations in non linear complex signal analysis. In: AGU FALL MEETING. American Geophysical Union, San Francisco, États-Unis (Dec 2011)