An efficient solution to sparse linear prediction analysis of speech

Vahid Khanagha, Khalid Daoudi

EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

An efficient solution to sparse linear prediction analysis of speech

Vahid Khanagha* and Khalid Daoudi

Abstract

We propose an efficient solution to the problem of sparse linear prediction analysis of the speech signal. Our method is based on minimization of a weighted l_2-norm of the prediction error. The weighting function is constructed such that less emphasis is given to the error around the points where we expect the largest prediction errors to occur (the glottal closure instants) and hence the resulting cost function approaches the ideal l_0-norm cost function for sparse residual recovery. We show that the efficient minimization of this objective function (by solving normal equations of linear least squares problem) provides enhanced sparsity level of residuals compared to the l_1-norm minimization approach which uses the computationally demanding convex optimization methods. Indeed, the computational complexity of the proposed method is roughly the same as the classic minimum variance linear prediction analysis approach. Moreover, to show a potential application of such sparse representation, we use the resulting linear prediction coefficients inside a multi-pulse synthesizer and show that the corresponding multi-pulse estimate of the excitation source results in slightly better synthesis quality when compared to the classical technique which uses the traditional non-sparse minimum variance synthesizer.

1 Introduction

Linear prediction (LP) analysis is a ubiquitous analysis technique in current speech technology. The basis of LP analysis is the source-filter production model of speech. For voiced sounds in particular, the filter is assumed to be an all-pole linear filter and the source is considered to be a semi-periodic impulse train which is zero most of the times, i.e., the source is a sparse time series. LP analysis results in the estimation of the all-pole filter parameters representing the spectral shape of the vocal tract. The accuracy of this estimation can be evaluated by observing the extent in which the residuals (the prediction error) of the corresponding prediction filter resemble the hypothesized source of excitation [1] (a perfect impulse train in case of voiced speech). However, it is shown in [1] that even when the vocal tract filter follows an actual all-pole model, this criterion of goodness is not fulfilled by the classical minimum variance predictor. Despite the theoretic physical significance, such sparse representation forms the basis for many applications in speech technology. For instance, a class of efficient parametric speech coders are based on the search for a sparse excitation sequence feeding the LP synthesizer [2].

It is argued in [3] that the reason behind the failure of the classical method in providing such sparse representation is that it relies on the minimization of l_2-norm of prediction error. It is known that the l_2-norm criterion is highly sensitive to the outliers [4], i.e., the points having considerably larger norms of error. Hence, l_2-norm error minimization favors solutions with many small non-zero entries rather than the sparse solutions having the fewest possible non-zero entries [4]. Hence, l_2-norm is not an appropriate objective function for the problems where sparseness constraints are incorporated. Indeed, the ideal solution for sparse residual recovery is to directly minimize the cardinality of this vector, i.e., the l_0-norm of prediction error which yields a combinatorial optimization problem. Instead, to alleviate the exaggerative effect of l_2-norm criterion at points with large norms of error, it is usual to consider the minimization of l_1-norm as it puts less emphasis on outliers. l_1-norm can be regarded as a convex relaxation of the l_0-norm and its minimization problem can be re-casted into a linear program and solved by convex programming techniques [5].

*Correspondence: vahid.khanagha@inria.fr
INRIA Bordeaux Sud-Ouest (GeoStat team), 200 Avenue de la Vielle Tour 33405 Talence, France

The l_1-norm minimization of residuals is already proven to be beneficial for speech processing [6-8]. In [6], the stability issue of l_1-norm linear programming is addressed and a method is introduced for both having an intrinsically stable solution as well as keeping the computational cost down. The approach is based the Burg method for autoregressive parameters estimation using the least absolute forward-backward error.

In [7], the authors have compared the Burg method with their l_1-norm minimization method using the modern interior points method and shown that the sparseness is not preserved with the Burg method. Later, they have proposed a re-weighted l_1-norm minimization approach in [8], to enhance the sparsity of the residuals and to overcome the mismatch between l_0-norm minimization and l_1-norm minimization while keeping the problem solvable with convex programming tools. Initially the l_1-norm minimization problem is solved using the interior points method and then the resulted residuals are used iteratively, to re-weight the l_1-norm objective function such that less weight is given to the points having larger residual norms. The optimization problem is thus iteratively approaching the solution for the ideal l_0-norm objective function. We also mention that, an interesting review is made in [9,10], on several solvers for the general problem of mixed l_p-l_0-norm minimization in the context of piece-wise constant function approximation, which indeed their adaptation to the problem of sparse linear prediction analysis can be beneficial (particularly the stepwise jump penalization algorithm, which is shown to be highly efficient and reliable in detection of sparse events).

In this article, we propose a new and efficient solution to sparse LP analysis which is based on weighting of the l_2-norm objective function so as to maintain the computational tractability of the final optimization problem and to avoid the computational burden of convex programming. The weighting function plays the most important role in our solution in maintaining the sparsity of the resulting residuals. We first extract from the speech signal itself, the points having the potential of attaining largest norms of residuals (the glottal closure instants) and then we construct the weighting function such that the prediction error is relaxed on these points. Consequently, the weighted l_2-norm objective function can be minimized by the solution of normal equations of liner least squares problem. We show that our closed-form solution provides better sparseness properties compared to the l_1-norm minimization using the interior points method. Also, to show the usefulness of such sparse representation, we use the resulting prediction coefficients inside a multi-pulse excitation (MPE) coder and we show that the corresponding multi-pulse excitation source provides slightly better synthesis quality compared to the estimated excitation of the classical minimum variance synthesizer.

The article is organized as follows. In Section 2, we provide the general formulation of the LP analysis problem. In Section 3, we briefly review previous studies on sparse LP analysis and the numerical motivations behind them. We present our efficient solution in Section 4. In Section 5, the experimental results are presented and finally in Section 6, we draw our conclusion and perspectives.

2 Problem formulation

The consideration of the vocal tract filter in the source-filter production model as an all-pole filter results in the well-known autoregressive model for the speech signal $x(n)$:

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

where a_k are the prediction coefficients, K is the order of prediction filter and $r(n)$ is the prediction error or the residual. In the ideal case, when the $\{a_k\}$ coefficients are perfectly estimated and the production mechanism verifies the all-pole assumption, the residual should resemble the hypothesized excitation source. In case of voiced speech, it should be a perfect semi-periodic impulse train which is zero most of the times, i.e., it is a sparse time series. The linear prediction analysis problem of a frame of length N can be written in the general matrix form as the l_p-norm minimization of the residual vector $\mathbf{r}$:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \parallel \mathbf{r} \parallel_p^p, \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a} \tag{2}$$

where $\mathbf{a}$ is a vector representing the set $\{a_k\}$ and

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{r} = \begin{bmatrix} r(N_1) \\ \vdots \\ r(N_2) \end{bmatrix},$$

$$\mathbf{X} = \begin{bmatrix} x(N_1-1) \cdots & x(N_1-K) \\ \vdots & \vdots \\ x(N_2-1) \cdots & x(N_2-K) \end{bmatrix}$$

and $N_1 = 1$ and $N_2 = N + K$ (For $n < 1$ and $n > N$, we put $x(n) = 0$). The l_p-norm is defined as $\parallel \mathbf{x} \parallel_p = (\sum_{k=1}^{N} x(k)^p)^{\frac{1}{p}}$. Depending on the choice of p in Equation 2, the estimated linear prediction coefficients and the resulting residuals would possess different properties.

3 Approaching the l_0-norm

The ideal solution to the LP analysis problem of Equation 2 so as to retrieve the sparse excitation source of voiced sounds, 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 [11]. As this problem is an N-P hard

optimization problem [8], its relaxed but more tractable versions ($p = 1, 2$) are the most widely used.

Setting $p = 2$ results in the classical minimum variance LP analysis problem. Although the latter suggests the highest computational efficiency, it is known that this solution can not provide the desired level of sparsity, even when the vocal tract filter is truly an all-pole filter [1]. It is known that l_2-norm has an exaggerative effect on the points having larger values of prediction error (the so-called outliers). Consequently, the minimizer puts much effort on forcing down the value of these outliers, with the cost of more non-zero elements. Hence, the resulting residuals are not as sparse as desired.

It is known that this exaggerative effect on the outliers is reduced with the use of l_1-norm and hence, its minimization could be a meliorative strategy w.r.t the minimum variance solution, in that the error on the outliers are less penalized [11]. The solution to the l_1-norm minimization is not as easy as the classical minimum variance LP analysis problem but it can be solved by recasting the minimization problem into a linear program [12] and then using convex optimization tools [5]. However, it is argued in [6] that linear programming l_1-norm minimization, suffers from stability and computational issues and instead, an efficient algorithm is introduced, based on a lattice filter structure in which the reflection coefficients are obtained using a Burg method with l_1 criterion and the robustness of the method is shown to be interesting for voiced sound analysis. However, it is shown in [7] that the l_1-norm Burg algorithms behaves somewhere in between the l_2-norm and the l_1-norm minimization. Instead, the authors have shown that enhanced sparsity level can be achieved using modern interior points method [5] of solving the linear program. They have shown interesting results of such analysis and have argued that the added computational burden is negligible considering the consequent simplifications (granted by such a sparse representation) in applications such as open and closed loop pitch analysis and algebraic excitation search.

An iteratively re-weighted l_1-norm minimization approach is consequently proposed by the same authors in [8] to enhance the sparsity of residuals, while keeping the problem solvable by convex techniques. The algorithm starts by plain l_1-norm minimization and then, iteratively, the resulting residuals are used to re-weight the l_1-norm cost function such that the points having larger residuals (outliers) are less penalized and the points having smaller residuals are penalized heavier. Hence, the optimizer encourages small values to become smaller while augmenting the amplitude of outliers [13].

The enhanced sparsity properties of the re-weighted l_1-norm solution compared to the l_1-norm minimization, and also the better performance of the l_1-norm criterion compared to l_2-norm criterion, can be explained numerically with the help of the graphical representation in Figure 1. There, the numerical effect of different residual values on l_p-norm cost functions is graphically depicted. It can be seen that the penalty on outliers is increasing with p. Indeed, as $p \rightarrow 0$ the penalty of the corresponding cost function on non-zero values approaches l_0-norm cost function (where any non-zero value is equally penalized and there is no penalization of larger values). This will force the minimization to include as many zeros as possible as their weight is zero. In case of the re-weighted l_1-norm solution [8], any residual is weighted by its inverse at each iteration and hence, the equal penalization of any non-zero value (as in l_0-norm criterion) is achieved. In other words, if a point has a very large (resp. very small) residual, it will be less (resp. much more) penalized in the next iteration and so, the sparsity is enhanced iteratively.

4 The weighted l_2-norm solution

We aim at developing an alternative and efficient optimization strategy which approximates the desired sparsity of the residuals. Our approach is based on the minimization of a weighted version of the l_2-norm criterion. The weighting function plays the key role in maintaining the sparsity of the residuals. Other than pure numerical motivations on de-emphasizing the exaggerative effect of l_2-norm on outliers (as discussed in Section 3), the design of this function is motivated by the physical production process of the speech signal. We extract from the speech signal itself, the points which are physically susceptible of attaining larger values of residuals and we construct the

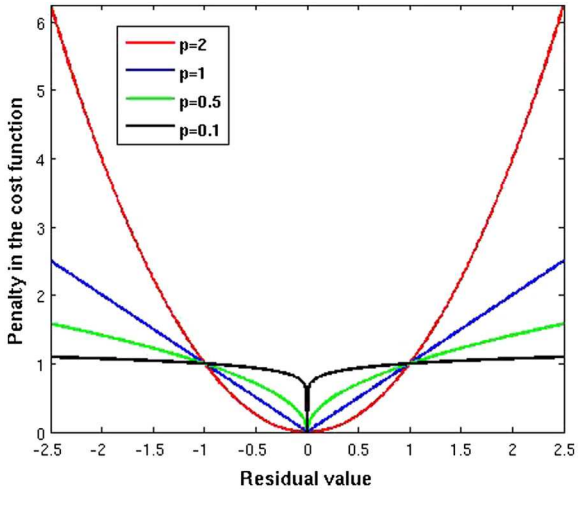

Figure 1 Graphical representation of different cost functions. Comparison between l_p-norm cost functions for $p \leq 2$. The "democratic" l_0-norm cost function is approached as $p \rightarrow 0$. The term "democratic" refers to the fact that l_0-norm weights all the nonzero coefficients equally [11].

weighting function such that the error at those outliers is less penalized.

The outliers of LP residuals have an important physical interpretation as their time-pattern follows the pitch period of the speech signal. In other words, they follow the physical excitation source of the vocal tract system, which is a sparse sequence of glottal pulses separated by the pitch period. Indeed, the impulse-like nature of this excitation source is reflected as effective discontinuities in the residual signal [14]: when no significant excitation is presented at the input of the vocal tract system, its output is resonating freely according to the hypothesized all-pole model, and hence, it is maximally predictable by the parameters of the filter. On the other hand, the predictability is minimized when the significant excitations takes place and hence, the output signal would be under the influence of both the excitation source and the vocal tract filter. Consequently, LP residual contains clear peaks (outliers) around the instants of significant excitations of the vocal tract system. Hence, if we have a-priori knowledge about these instants we can use this knowledge to impose constraints on the LP analysis problem, so as to relax the prediction error at those points. By doing so, we ensure that if any enhancement is achieved in the sparsity level of residuals, it also corresponds to the physical sparse source of excitation.

The instants of significant excitations of vocal tract are called the Glottal closure instants (GCI) [14]. The detection of GCIs has gained significant attention recently as it founds many interesting applications in pitch-synchronous speech analysis. Many methods are developed for GCI detection in adverse environment (a comparative study is provided in [15]) and the physical significance of the detected GCIs is validated by comparing them to the electro Glotto graph signal. In this article, we use the recent robust SEDREAMS algorithm [15][a]. The weighting function is then constructed such that less emphasize is given to the GCI points, and hence the exaggerative effect on the outliers of the residuals is canceled. We can now proceed to formalize the proposed solution.

4.1 Optimization algorithm

We opt for l_2-norm cost function to preserve computational efficiency and then we cope with its exaggerative effect on outliers, by careful down-weighting of the cost function at those points. Formally, we define following optimization problem for the recovery of sparse residuals:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \quad \sum_{k=1}^{N} w(k)(r(k)^2) \tag{3}$$

where $w(\cdot)$ is the weighting function. Once $w(\cdot)$ is properly defined, the solution to Equation 3 is straight-forward. Indeed, setting the derivative of the cost function to zero results in a set of normal equations which can be solved as in the classical l_2-norm approach:

$$\hat{\mathbf{a}} = \mathbf{R}^{-1}\mathbf{p} \tag{4}$$

while in our case, $\mathbf{R} = (\mathbf{W} \odot \mathbf{X})\mathbf{X}^T$, $\mathbf{p} = \mathbf{w} \odot (\mathbf{X}^T\mathbf{x})$, $\odot$ denotes the element-wise product of the two matrices and:

$$\mathbf{w} = \begin{bmatrix} w(N_1) \\ \vdots \\ w(N_2) \end{bmatrix}, \mathbf{W} = \begin{bmatrix} w(N_1) & \cdots & w(N_1) \\ \vdots & & \vdots \\ w(N_2) & \cdots & w(N_2) \end{bmatrix}$$

It is interesting to mention that our experiments show that as long as the smoothness of the $w(\cdot)$ is maintained the stability of the solution is preserved. Indeed, the special form of the input vector $\mathbf{X}$ in Equation 2, is the one used in autocorrelation formulation of LP analysis using l_2-norm minimization. It is proven that autocorrelation formulation always results in a minimum-phase estimate of the all-pole filter, even if the real vocal tract filter is not minimum phase [1]. As our formulation is similar to the autocorrelation formulation, we can fairly expect the same behavior (though we don't have a theoretical proof). This is indeed beneficial, as having a non-minimum phase spectral estimate results in saturations during synthesis applications. Our experiments show that such saturation indeed never happens. This is an interesting advantage of our method compared to l_1-norm minimization methods which do not guaranty a minimum phase solution, unless if additional constraints are imposed to the problem [7].

4.2 The weighting function

The weighting function is expected to provide the lowest weights at the GCI points and to give equal weights (of one) 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 function. We thus define the final weighting function as:

$$w(n) = 1 - \sum_{k=1}^{N_{\text{gci}}} g(n - T_k) \tag{5}$$

where $T_k, k = 1 \ldots N_{\text{gci}}$ 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. Figure 2 shows a frame of voiced sound along with the GCI points detected by the SEDREAMS algorithm and the weighting function of Equation 5. It can be seen that this weighting function puts the lowest weights around the GCI locations (i.e., the expected outliers) and equally weights the remaining points. Numerically speaking, the minimizer is free to pick

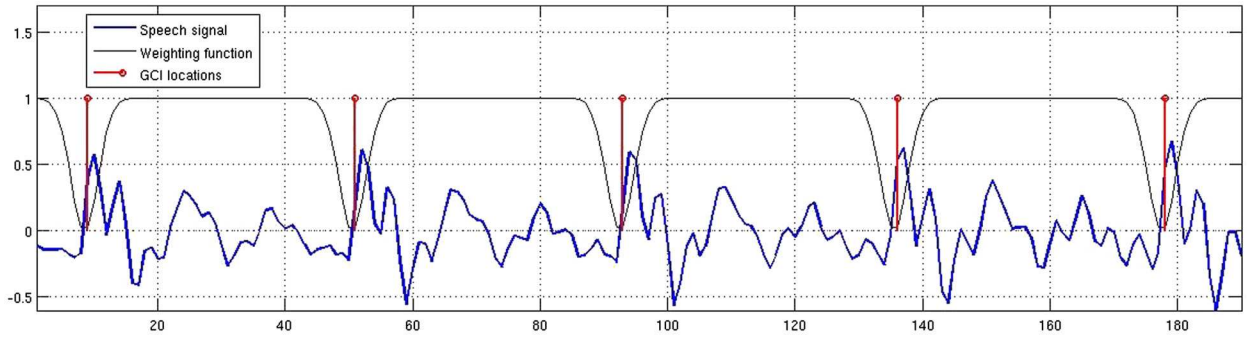

Figure 2 The weighting function A frame of a voiced sound along with the detected GCI locations and the constructed weighting function (with $\sigma = 50$ and $\kappa = 1$).

the largest residual values for the outliers and it concentrates on minimizing the error on the remaining points (hence the sparsity is granted as explained in Section 3). This can also be explained with regard to physical production mechanism of the speech signal: as the coupling of excitation source and vocal tract filter is maximized on GCIs, such weighting function assists the minimizer to exclude the points on which the coupling is maximized and concentrate its effort on speech samples where the source contribution is minimized. Such decoupling is investigated in the context of Glottal volume velocity estimation by closed phase inverse filtering techniques [16]. There, the whole time interval on which the glottis is expected to be open is localized and discarded from the analysis frame. Consequently, these methods require the availability of both GCI and Glottal opening instants (GOI). However, the determination of GOIs is much more difficult than GCI detection [16]. Moreover, as the analysis window is strictly limited to the closed phase [17], another practical issue may arise: this time-frame might be too short (for high-pitched voices for instance) such that the analysis becomes ineffective [16].

5 Experimental results

We first show the ability of our approach in retrieving sparse residuals for stationary voiced signals and also we show that it can provide a better estimation of the all-pole vocal-tract filter parameters. We then show how such sparse modeling can enhance the performance of a multi-pulse excitation estimation. 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 [18].

5.1 Sparsity of residuals for voiced sounds

We compare the performance of our weighted-l_2-norm solution with that of the classic l_2-norm minimization and also the l_1-norm minimization via convex programming. For minimization of the l_1-norm, we use the publicly available l_1-magic toolbox [12] which uses the primal-dual interior points optimization [5]. Figure 3 shows the residuals obtained for 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 our weighted-l_2-norm solution shows better sparsity properties compared to the l_1-norm minimization: in the former, each positive peak of residuals is followed by a single negative peak (of almost the same amplitude) while for the latter, any positive peak is surrounded by two negative peaks of smaller (but yet significant) values.

This comparison can be further formalized by using a quantitative measure of sparsity. There exists plenty of such measures on which a review is provided in [19]. Among them, we use the kurtosis, as it satisfies three of the most important properties that are intuitively expected from a measure of sparsity: scale invariance, rising tide and Robin Hood [19]. 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 three optimization strategies, averaged over 20 randomly selected sentences of both male and female speakers taken from TIMIT database. From the table, it is clear the our method achieves the highest level of sparsity as it obtains the highest values of the kurtosis.

5.2 Estimation of the all-pole vocal-tract filter

We also investigate the ability of our method in estimation of the all-pole filter parameters. To do so, we generate a synthetic speech signal by exciting a known all-pole system with a periodic sequence of impulses (at known locations). We then estimate these parameters from the synthetic signal by LP analysis using our method and the classical l_2-norm method. Figure 4 shows the frequency response of the resulting estimates along with the

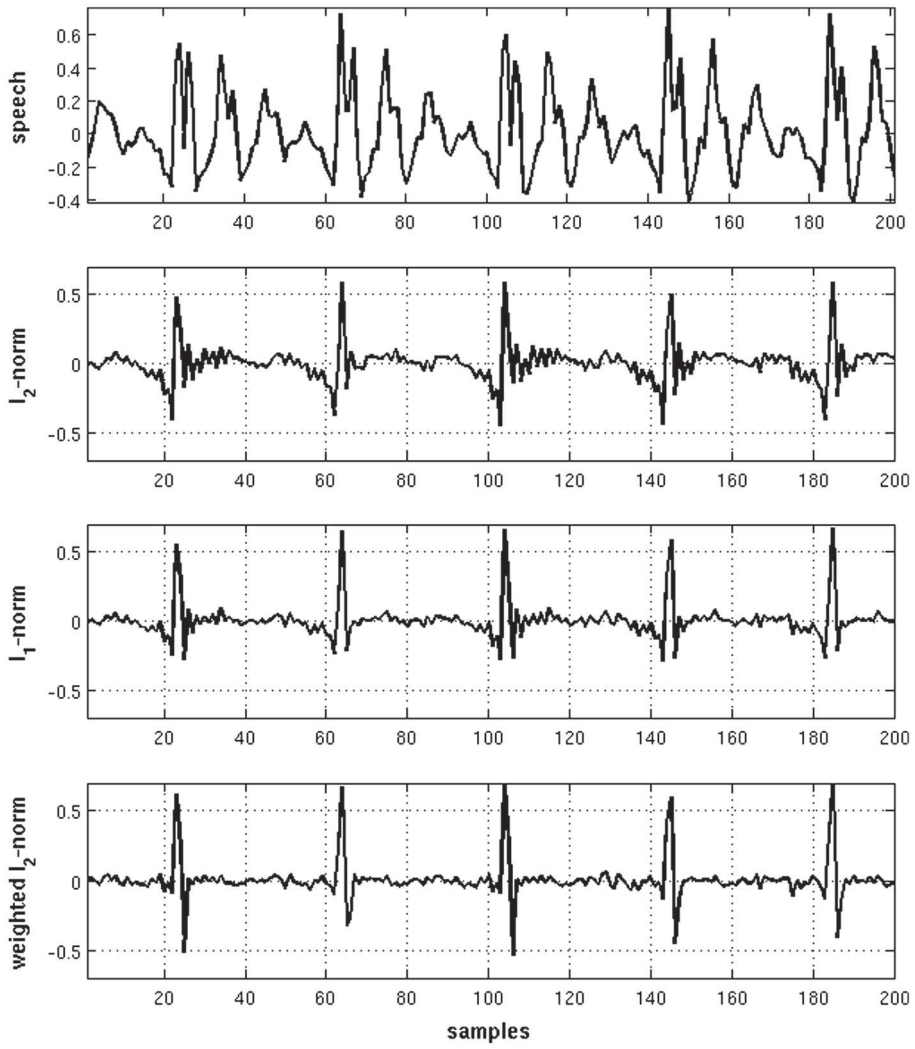

Figure 3 Comparison of the sparsity of residuals. The residuals of the LP analysis obtained from different optimization strategies. The prediction order is $K = 13$ and the frame length is $N = 160$.

frequency-domain representation of the synthetic excitation source. It can be seen that for the l_2-norm minimizer, there is a clear shift in the peaks of the estimated filter towards the harmonics of the excitation source. Specifically, the first spectral peak is shifted toward the forth harmonic of the excitation source. Indeed, the effort of l_2-norm minimizer in reducing great errors (the outliers due to the excitation source), has caused the estimated filter to be influenced by the excitation source. However, our weighted-l_2-norm minimization makes a very well estimation of the original all-pole filter and there is no shift in the spectral peaks. This shows that our method effectively decouples the contributions of the excitation source and the all-pole filter (as the source contribution is de-emphasized by the weighting function).

Table 1 Comparison of the sparsity levels

Method	kurtosis on the whole sentence	kurtosis on voiced parts
l_2-norm minimization	51.7	39.7
l_1-norm minimization	81.9	65.9
weighted-l_2-norm minimization	86.6	69.7

Quantitative comparison of the level of sparsity of different LP analysis strategies.

5.3 Multi-pulse excitation estimation

The sparseness of the excitation source is a fundamental assumption in the framework of linear predictive coding (LPC) where the speech is synthesized by feeding the estimated all-pole filter by an estimate of the excitation

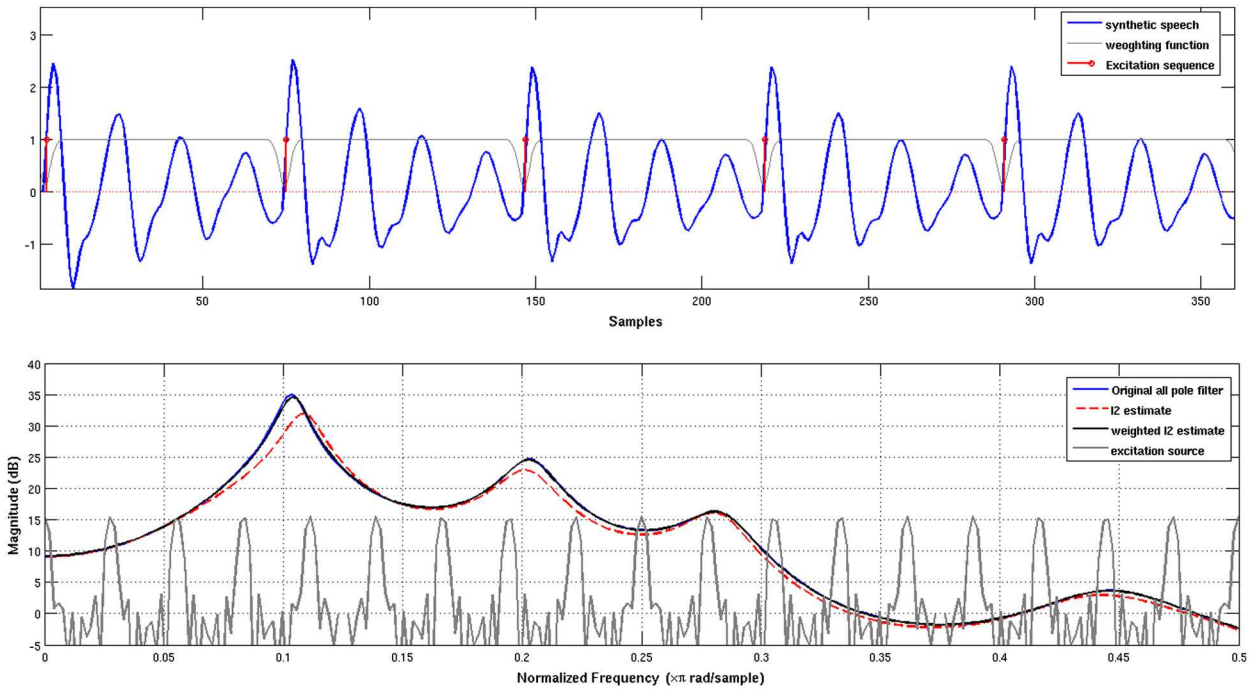

Figure 4 Estimation of the all-pole filter. (top) synthetic speech signal, (bottom) frequency response of the filters obtained with l_2-norm and weighted-l_2-norm minimization (prediction order $K = 13$). Note that only the first half of the frequency axis is shown so as to enhance the presentation.

source. The coding gain is achieved by considering a sparse representation for the excitation source. In the popular multi-pulse excitation (MPE) method [20,21], the synthesis filter is estimated through the classic l_2-norm minimization and then a sparse multi-pulse excitation sequence is extracted through an iterative Analysis-by-Synthesis procedure. However, as discussed in previous sections this synthesizer is not intrinsically a sparse one. Hence, it would be logical to expect that the employment of an intrinsically sparse synthesis filter, such as the one developed in this article, would enhance the quality of the synthesized speech using the corresponding multi-pulse estimate. Consequently, we compare the performance of the classical MPE synthesizer which uses minimum variance LPC synthesizer with the one whose synthesizer is obtained trough our weighted l_2-norm minimization procedure. We emphasize that we follow exactly the same procedure for estimation of multipulse coders for both synthesizers, as in the classical MPE implementation in [21] (iterative minimization of perceptually weighted error of reconstruction).

We have tried to follow the same experimental protocol as in [22]. That is, we evaluate our method using about 1 h of clean speech signal randomly chosen from the TIMIT database (re-sampled to 8 kHz) uttered by speakers of different genders, accents and ages which provides enough diversity in the characteristics of the analyzed signals. Thirteen prediction coefficients are computed for frames of 20 ms ($N = 160$) and the search for the multi-pulse sequence (10 pulses per frame) is performed as explained in [21]. We evaluate the quality of reconstructed speech in terms of SNR and the PESQ measure [23] which provides a score of perceptual quality in the range of 1 (the worst quality) to 5 (the best quality). The results are shown in Table 2, which shows that our method achieves slightly higher coding quality than the classical MPE synthesizer.

Finally, we emphasize that the superior performance of our weighted-l_2-norm solution in retrieving sparse residuals in Section 5.2, plus the slight improvement of the coding quality in Section 5.3, was achieved with roughly the same computational complexity as the classical l_2-norm minimization (if we neglect the computational burden of the GCI detector). This is a great advantage compared to the computationally demanding l_1-norm minimization

Table 2 Multi-pulse excitation coding

Method	PESQ	SNR
MPE + l_2-norm	3.3	9.5 dB
MPE + weighted-l_2-norm	3.4	10.2 dB

The quality of Multi-pulse excitation coding using two different synthesizer filters. The multi-pulse excitation source of MPE coder is constructed by taking 10 pulses per 20 ms.

via convex programming (as in [7] or in [8] where multiple re-weighted l_1-norm problems are solved) which also suffers from instability issues. Moreover, another important feature of our solution is that, during the coding experiment we observed that by using the Gaussian shape for the weighting function, the solution is always stable and it does not meet the instability issues as l_1-norm minimization.

6 Conclusion

We introduced a simple and efficient solution to the problem of sparse residual recovery of the speech signal. Our approach is based on minimization of weighted l_2-norm of the residuals. The l_2-norm was used to preserve the simplicity and the efficiency of the solution, while the weighting function was designed to circumvent l_2-norm's exaggerative effect on larger residuals (the outliers). This is done by de-emphasizing the error on the GCIs where, by considering the physical production mechanism of the speech, we expect the outliers to occur. We showed that our methodology provides better sparsity properties compared to the complex and computationally demanding l_1-norm minimization via linear programming. Moreover, the method is interestingly immune to instability problems as opposed to l_1-norm minimization. Also, we showed that such intrinsically sparse representation can result in slightly better synthesis quality by sparse multi-pulse excitation of the synthesis filter in MPE-coding framework. The performance, the efficiency and the stability of the proposed solution show a promising potential in speech processing as its application can be further investigated in a variety of applications in the general framework of speech synthesis. This would be the subject of our future communications.

Endnote

[a]We opted for the SEDREAMS so as to benefit from its proven reliability, in order to focus on proof-of-concept.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The main contribution of this article is the introduction of a simple, efficient and stable solution for retrieving sparse residuals of the linear prediction analysis of the speech signal. Our solution is based on minimization of a weighted version of the l_2-norm prediction error. The l_2-norm was used to provide a simple and efficient solution, while the weighting function was designed to circumvent l_2-norm's exaggerative effect on the outliers of linear prediction error. The weighting function is thus designed such that less weight is given to the instants of significant excitation of vocal tract system (the glottal closure instants), as we expect the prediction error to be maximized at these points. Our method shows superior sparsity properties compared to the classical l_2-norm and also the modern l_1-norm minimization techniques. Our method has roughly the same computational complexity as the classical minimum variance solution (contrary to the computationally demanding l_1-norm minimization using convex programming) and it does not suffer from the instability issues as l_1-norm minimization. Moreover, we showed that our intrinsically sparse representation can slightly enhance the synthesis quality by sparse multi-pulse excitation of the synthesis filter in MPE-coding framework. Both authors read and approved the final manuscript.

Acknowledgements

Vahid Khanagha is funded by the INRIA CORDIS doctoral program.

Received: 15 June 2012 Accepted: 27 October 2012
Published: 22 January 2013

References

1. TF Quatieri, *Discret-Time Speech Signal Processing Principles and Practice*. (Prentice-Hall, Upper Saddle River, New Jersey, USA, 2001)
2. WC CHU, *Speech coding algorithms: foundation and evolution of standardized coders*. (Wiley-Interscience, John Wiley & Sons, Inc., Hoboken, New Jersey, USA, 2003)
3. D Giacobello, Sparsity in Linear Predictive Coding of Speech. PhD thesis. Multimedia Information and Signal Processing, Department of Electronic Systems, Aalborg University (2010)
4. D Meng, Q Zhao, Z Xu, Improved robustness of sparse PCA by L1-norm maximization. Pattern Recogn. Elsevier. **45**, 487–497 (2012)
5. S Boyd, L Vandenberghe, *Convex Optimization*. (Cambridge University Press, Shaftesbury Road, Cambridge, United Kingdom, 2004)
6. E Denoel, JP Solvay, Linear prediction of speech with a least absolute error criterion. IEEE Trans. Acoustics Speech Signal Process. **33**, 1397–1403 (1985)
7. D Giacobello, MG Christensen, J Dahl, SH Jensen, M Moonen, in *Proceedings of the INTERSPEECH*. Sparse linear predictors for speech processing (Brisbane, Australia, 2009), pp. 353-1356
8. D Giacobello, MG Christensen, MN Murthi, SH Jensen, M Moonen, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization (Dallas, Texas, USA, 2010), pp. 4650-4653
9. MA Little, N Jones, Generalized methods and solvers for noise removal from piecewise constant signals: Part I–Background theory. Proceedings of the Royal Society A. **467**(2135), 3088–3114 (2011)
10. MA Little, N Jones, Generalized methods and solvers for noise removal from piecewise constant signals: Part II–New methods. Proceedings of the Royal Society A. **467**(2135), 3115–3140 (2011)
11. EJ Candès, MB Wakin, Enhancing sparsity by reweighted l1 minimization. J. Fourier Anal. Appl. **14**, 877–905 (2008)
12. E Candès, J Romberg, '1-MAGIC: Recovery of sparse signals via convex programming. California Institute of Technology, Pasadena. 24 (2005)
13. D Giacobello, MG Christensen, MN Murth, SH Jensen, Moonen Marc F, Sparselinear prediction, itsapplicationstospeechprocessing Trans, IEEE Audio Speech Lang. Process. **20**, 1644–1657 (2012)
14. K Murty, B Yegnanarayana, Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. **16**, 1602–1613 (2008)
15. T Drugman, M Thomas, J Gudnason, P Naylor, T Dutoit, Detection of glottal closure instants from speech signals: a quantitative review. IEEE Trans. Audio Speech Lang. Process. **20**(3), 994–1006 (2012)
16. M Thomas, J Gudnason, P Naylor, Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm. IEEE Trans. Audio Speech Lang. Process. **20**(1), 82–97 (2012)
17. D Wong, J Markel, JA Gray, Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Trans. Acoustics Speech Signal Process. **27**(4), 350–355 (1979)
18. JS Garofolo, LF Lamel, WM Fisher, JG Fiscus, DS Pallett, NL Dahlgren, V Zue, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. Tech. rep., U.S. Dept. of Commerce, NIST*. (MD, Gaithersburg, 1993)
19. N Hurley, S Rickard, Comparing measures of sparsity. IEEE Trans. Inf. Theory. **55**, 4723–4740 (2009)
20. B Atal, J Remde, vol. 7. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), (1982), pp. 614–617
21. S Singhal, B Atal, Amplitude optimization and pitch prediction in multipulse coders. IEEE Trans. Acoustics Speech Signal Process. **37**, 317–327 (1989)

22. D Giacobello, M Christensen, M Murthi, S Jensen, M Moonen, Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction. IEEE Signal Process. Lett. **17**(1), 103–106 (2010)
23. International Telecommunication Union: ITU-T Recommendation p. 862: Perceptual evaluation of speech quality (PESQ). an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs

doi:10.1186/1687-4722-2013-3
Cite this article as: Khanagha and Daoudi: **An efficient solution to sparse linear prediction analysis of speech.** *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:3.