A GREEDY APPROACH TO LINEAR PREDICTION WITH SPARSE RESIDUALS

Yuantao Gu

State Key Laboratory on Microwave and Digital Communications
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, CHINA
gyt@tsinghua.edu.cn

ABSTRACT

This paper focuses on the problem of Linear Prediction (LP) constrained by sparse residuals. After reformulating the problem to finding the largest linear correlated strict subset in a given vector set, a greedy method is proposed to determine the support of the sparse residuals iteratively by testing each entry with respective temporary prediction error. The greedy method is then simplified to reduce computational cost. Compared with reference algorithms and conventional LP model, the proposed methods are tested in the speech coding scenario. Experiment results demonstrate that the proposed greedy methods work well and suggest that LP with sparse residuals provides accurate estimation, and is much practical in the scenarios that more bits are allocated for coding residuals.

***Index Terms*—** Linear prediction, sparsity, sparse residual, greedy algorithm, speech coding, quantization, the largest linear correlated strict subset.

1. INTRODUCTION

Linear Prediction (LP) model [1, 2] has been well applied in numerous speech applications [3]. For example, many speech coding standards are developed based on Code-Excited Linear Prediction (CELP) [4, 5]. In such applications, the speech signal is assumed to be modeled as the output of an all-pole filter. It is further supposed that the all-pole filter varies rather slowly in short time period and the excitation is a pulse sequence, a pseudo-random noise, or their combinations. Consequently, Least Square (LS) minimization in the excitation (or the residual) domain is adopted to estimate the filter coefficients first, and then the excitation or prediction residual is produced.

At the decoder side, the synthetic speech is generated from the all-pole filter excited by the LP residuals, while the quantized filter coefficients and the residuals are received from the encoder side. To yield synthetic speech with high quality and to avoid the nonlinear effect of quantization, the quantization is performed in a *closed-loop* synthesis approach of searching in a codebook by LS minimization in the perceptually weighted speech domain. In other words, LP based speech coding is conducted by finding an all-pole filter and a pulse sequence in respective codebook that the decoder may use them to synthesize a signal sounds like the original speech.

Since 1980s, there had always been an attempt to apply different criteria other than LS minimization to estimate the all-pole filter coefficients more accurately or robustly[6, 7, 8, 9, 10, 11, 12, 13]. Most of them are based on l_p norm constraint, where p is equal to or less than 1, to generate sparse prediction residuals. Accompanying the popular Compressive Sensing (CS) [14] and the hot topic of sparse recovery, LP with sparse constraint has attracted much attention in recent years.

In this paper, different from available algorithms, the greedy approach is adopted for the first time to solve the problem of LP with sparse residuals (LPSR). In section 2, the sparse constrained LP problems are introduced through a brief review of their evolutions. In section 3, two greedy algorithms are proposed to solve LPSR problem, which is formulated to a problem of finding the largest linear correlated strict subset in given vectors. In section 4 the sparse constrained LP problem is readily casted into the CS model and can be solved by standard sparse signal recovery algorithm. The experiments are conducted in section 5. The proposed algorithms are verified to behave well in the real speech scenario. Furthermore, it is found that the sparse constrained LP works better than conventional LP in moderate rate speech coding scenario, where more bits could be allocated to represent residuals. The conclusion goes in section 6.

2. LINEAR PREDICTION WITH SPARSE CONSTRAINTS

Let's define the LP model as

$$s_n = \sum_{k=1}^{F} a_k s_{n-k} + r_n, \tag{1}$$

where s_n, r_n, a_k, and F denote the speech signal, the prediction residual, and the all-pole filter coefficients and order, respectively. The first milestone of linear prediction by l_1 minimization, i.e.

$$\{\hat{a}_k\} = \arg\min \sum_n |r_n| \tag{2}$$

is placed by Denoel and Solvay before 1985 [6]. Their motivation is to provide a robust LP estimator, as well as to preserve the pulse-like entries in the residual signal as large prediction errors. They adopted the lattice filter structure and modified the Burg algorithm with l_1 criterion to generate a stable all-pole filter. The algorithm was further improved by means of orthogonal transformation [7]. Nearly at the same time, Nammone, Weng, and Gay [8] studied the same model and solved (2) by simplex method of linear programming. Consequently, Lansford and Yarlagadda described some experimental results using $l_p(1 \le p \le 3)$ normed models in the speech coding scenario[9]. They solve the problem by the residual steepest descent algorithm based on their previous work on l_p deconvolution [15].

Besides the l_p constraint approach, Lee tried to minimize the sum of appropriately weighted residuals [10], i.e.

$$\{\hat{a}_k\} = \arg\min \sum_n \rho(r_n), \tag{3}$$

where the weight function $\rho(\cdot)$ is selected to give more weight to the smaller residuals while less weight to the large residuals, which are in small portion. Their results demonstrated the new formulation provided a more efficient and less biased estimate for the prediction coefficients, compared to conventional LP.

Later in the 1990s, Namba, Kamata, and Ishida [16] further studied (2) and demonstrated experimentally that the speech spectrum envelop derived from l_1 estimator is changing smoothly with regard to time, while that from l_2 estimator changes sharply. The second milestone was placed by Murthi and Rao in 1998 [11]. They tried to generate sparse residues explicitly suiting for codebook excitation with an error criterion

$$\{\hat{a}_k\} = \arg\min \lim_{p\to 0} \sum_n |r_n|^p \tag{4}$$

and solved by an Iteratively Reweighted Least Squares algorithm[12]. This is the first appearance of l_p $(p < 1)$ norm minimization in LP model. It can be readily accepted that the sparsity of the residual signal is put at the first place in (4) by approximate l_0 norm minimization.

After the rising and flourishing of Compressive Sensing and sparse recovery, Giacobello, Christensen, Murthi, Jensen, and Moonen did a series of solid works [17, 18, 13] on this area and named Sparse Linear Prediction (SLP) as

$$\{\hat{a}_k\} = \arg\min |\{r_n\}| + \gamma|\{a_k\}|, \tag{5}$$

where γ is a factor to balance the two parts. Please notice that both short term (formant filter) and long term (pitch filter) are combined together as the all-pole filter model, whose inverse is obviously a sparse filter. They solved the problem by iteratively reweighted l_1-norm minimization of the residual and the coefficients. Furthermore, they have successfully built several speech codecs, which exceed some commercial speech coding standard in various tests [13].

I will end this brief review by recalling some more early references. Least squares estimation provides a maximum likelihood (ML) estimate in the presence of Gaussian noise, while l_1 minimization is the ML estimate when the noise is Laplacian. Furthermore, maximum *a posteriori* (MAP) estimation can be seen as a regularization of ML estimate. Lim and Oppenheim had established a theoretical foundation for the estimation of an all-pole model parameters by MAP criterion in the scenario of speech degraded by background noise[19], which is the earliest work I found that contained a well-defined LP model with rather general error criteria.

3. THE PROPOSED GREEDY ALGORITHMS

This work will only focus on the sparsity in residual domain, other than the prediction coefficients. LP model of (1) can be rewritten in a matrix multiplication formulation

$$\begin{bmatrix} s_n & s_{n-1} & \cdots & s_{n-F} \\ s_{n-1} & s_{n-2} & \cdots & s_{n-F-1} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n-L+1} & s_{n-L} & \cdots & s_{n-L-F+1} \end{bmatrix} \begin{bmatrix} 1 \\ -a_1 \\ \vdots \\ -a_F \end{bmatrix} = \begin{bmatrix} r_n \\ r_{n-1} \\ \vdots \\ r_{n-L+1} \end{bmatrix}, \tag{6}$$

where L denotes the frame length. LP problem is to find a coefficient vector $\mathbf{a} = [1, -a_1, -a_2, \cdots, -a_F]^{\mathrm{T}}$, which makes the residual vector $\mathbf{r} = [r_n, r_{n-1}, \cdots, r_{n-L+1}]^{\mathrm{T}}$ satisfy certain optimization constraint.

Particularly, if $\mathbf{r}$ is supposed to be a nearly sparse vector to play the role of a multi-pulse-like excitation in speech synthesis, one may specify the model in (6) as

$$\mathbf{Sa} = \mathbf{r}^{\mathrm{s}} + \mathbf{r}^{\mathrm{n}}, \tag{7}$$

where $\mathbf{S}$ denotes the matrix contains speech samples on the left side of (6), $\mathbf{r}^{\mathrm{s}}$ and $\mathbf{r}^{\mathrm{n}}$ are, respectively, the exact sparse component and dense noise component of the residual. Consequently, LPSR is formed as an optimization problem,

$$\hat{\mathbf{a}} = \arg\min \|\mathbf{Sa} - \mathbf{r}^{\mathrm{s}}\|_2, \quad \text{s.t.} \quad \|\mathbf{r}^{\mathrm{s}}\|_0 = K, \tag{8}$$

where K is a predefined number of non-zero entries.

Furthermore, two assumptions are adopted to clearly define the sparsity of the exact sparse residual vector, which will be used in the following analysis. *Assumption 1:* the magnitudes of all non-zero entries in the exact sparse residual vector are larger than those in the noisy residual vector. *Assumption 2:* the number of non-zero entries in the exact sparse residual vector is much less than the vector length.

In this work, a first-greedy-then-LS approach is adopted to solve (8). Denoting the support set of the exact sparse residual by U, one may define those rows in speech matrix $\mathbf{S}$ and entries in noisy residual $\mathbf{r}^{\mathrm{n}}$ corresponding to the index set U by $\mathbf{S}_U$ and $\mathbf{r}^{\mathrm{n}}_U$, respectively. If one has an estimate of the support of sparse residual, i.e. $\hat{U}$, then the prediction coefficients can be calculated by LS minimization subject to those rows, which are not in $\hat{U}$, in (7),

$$\hat{\mathbf{a}} = \arg\min \left\|\mathbf{r}^{\mathrm{n}}_{\hat{U}_{\mathrm{C}}}\right\|_2 = \arg\min \left\|\mathbf{S}_{\hat{U}_{\mathrm{C}}}\mathbf{a}\right\|_2, \tag{9}$$

where $\hat{U}_{\mathrm{C}} = \{1, 2, \cdots, L\} \backslash \hat{U}$. The optimality of (9) can by readily explained from another point of view as follows. There are L linear measurements of F unknowns, where $L \gg F$. If one further knows some measurements containing strong noises which correspond to the support set $\hat{U}$, then the optimal estimate should be the LS solution from the rest of measurements, which are less noisy. The above analysis also suggests that LPSR could provide a more accurate estimate than conventional LP.

Consequently, the essential of LPSR is to determine the support set of the exact sparse residual. By revisiting (7) and utilizing Assumption 1, one has

$$\left|\mathbf{s}_i^{\mathrm{T}}\mathbf{a}\right| \gg \left|\mathbf{s}_j^{\mathrm{T}}\mathbf{a}\right|, \quad \forall i \in U, \quad \forall j \in U_{\mathrm{C}} \tag{10}$$

where $\mathbf{s}_i^{\mathrm{T}} = [s_{n-i+1}, s_{n-i}, \cdots, s_{n-F-i+1}]$ denotes the ith row in speech matrix $\mathbf{S}$. Focusing on the two vector sets $\mathcal{S}_U = \{\mathbf{s}_i\}_{i\in U}$ and $\mathcal{S}_{UC} = \{\mathbf{s}_i\}_{i\in U_{\mathrm{C}}}$, one may readily accept from (10) that, comparing to those in the former set, the vectors in the latter one *nearly* locate in a hyperplane, to which a non-zero vector $\mathbf{a}$ is vertical. Specifically, if there is no noisy residual in LPSR model, i.e. $\left|\mathbf{s}_j^{\mathrm{T}}\mathbf{a}\right| = 0, \forall j \in U_{\mathrm{C}}$, the prediction vector $\mathbf{a}$ is exactly vertical to the hyperplane that $\mathcal{S}_{U_{\mathrm{C}}}$ composes.

Let's now adopt Assumption 2, which means that the cardinality of $\mathcal{S}_U$ is rather small comparing to that of $\mathcal{S}_{U_{\mathrm{C}}}$. Then it is glad to recognize that almost all vectors in $\mathcal{S} = \{\mathbf{s}_i\}$ compose a hyperplane, and the task is to pick those far from the mentioned plane. Please refer to Fig. 1 for a visualization.

Therefore, the problem is reformulated as: *for a given vector set $\mathcal{S}$, one tries to find a non-zero vector $\mathbf{a}$ that is orthogonal to most of the vectors in $\mathcal{S}$, or equivalently, one tries to find the largest linear correlated strict subset $\mathcal{S}_{U_{\mathrm{C}}} \subset \mathcal{S}$.*

Based on the above analysis, a greedy algorithm to find the support set of the sparse residual is proposed in Table 1. The algorithm

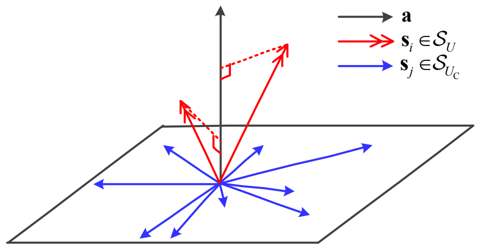

Fig. 1. A geometric interpretation to LPSR problem.

Table 1. The Proposed Greedy Algorithm

Input: $\mathbf{s}_j (j = 1, 2, \cdots, L)$, K
Initialization: $\hat{U}_{\mathrm{C}}^{(1)} = \{1, 2, \cdots, L\}$
Output: $\hat{U}_{\mathrm{C}}^{(K)}$
For: $i = 1, 2, \cdots, K$
Evaluate:
$\varepsilon_k^{(i)} = \min_{\mathbf{a}} \sum_{j \in \hat{U}_{\mathrm{C}}^{(i)} \backslash k} \|\mathbf{s}_j^{\mathrm{T}} \mathbf{a}\|_2, \quad \forall k \in \hat{U}_{\mathrm{C}}^{(i)}$
Choose:
$t^{(i)} = \arg\min_{k \in \hat{U}_{\mathrm{C}}^{(i)}} \varepsilon_k^{(i)}$
Remove:
$\hat{U}_{\mathrm{C}}^{(i+1)} = \hat{U}_{\mathrm{C}}^{(i)} \backslash t^{(i)}$
End

iteratively estimate the complementary set $\hat{U}_{\mathrm{C}}$. In the ith iteration, the algorithm eliminates the index $t^{(i)}$, the vector of which is the most unlikely to lie in the hyperplane, from the current complementary set $\hat{U}_{\mathrm{C}}^{(i)}$. The likelihood comes from the following observation. If $k \in U$, the temporary complementary set $\mathcal{S}_{\hat{U}_{\mathrm{C}}^{(i)} \backslash k}$ shall behave more like a hyperplane than $\mathcal{S}_{\hat{U}_{\mathrm{C}}^{(i)}}$, one could be easier to find a prediction vector $\mathbf{a}$ nearly vertical to all vectors in the temporary set. In other words, the prediction error should be smaller than the case that $k \in U_{\mathrm{C}}$ is temporarily removed. Therefore the prediction error $\varepsilon_k^{(i)}$ for a specific removed k is selected to evaluate the likelihood of $k \in U$: the smaller $\varepsilon_k^{(i)}$ is, the more likely k is to be in U. Please refer to Table 1 for detailed description.

The computational cost of the proposed algorithm is $\mathcal{O}(KLF^2 + \alpha KF^3)$, where α is a factor. Please notice that there are L similar $F \times F$ matrix need be inverted in every *Evaluate* step, where Sherman-Morrison formula [20] is adopted to reduce the multiplication times from $\mathcal{O}(LF^3)$ to $\mathcal{O}(LF^2 + \alpha F^3)$.

To further reduce the computational cost, a simplified greedy algorithm is proposed in Table 2 to balance the computation cost and performance. In this algorithm, the candidates of U are selected at once other than in K iterations, based on the likelihood defined exactly the same as in Table 1. Consequently, the computational cost is reduced to $\mathcal{O}(LF^2 + \alpha F^3)$.

Relation to prior work: As far as I know, this is the first appearance of greedy approach to solve LPSR problem. Simulations verifies that its performance exceeds the reference algorithms.

4. LPSR AND COMPRESSIVE SENSING

The renaissance of LPSR is closely relevant to the boom of CS and sparsity related topics. The connection between LPSR and CS has

Table 2. The Simplified Greedy Algorithm

Input: $\mathbf{s}_j, (j = 1, 2, \cdots, L)$, K
Initialization: $\hat{U}_{\mathrm{C}} = \{1, 2, \cdots, L\}$
Output: $\hat{U}_{\mathrm{C}}$
Execute:
Evaluate:
$\varepsilon_k = \min_{\mathbf{a}} \sum_{j \in \hat{U}_{\mathrm{C}} \backslash k} \|\mathbf{s}_j^{\mathrm{T}} \mathbf{a}\|_2, \quad \forall k \in \hat{U}_{\mathrm{C}}$
Choose:
$\hat{U}_{\mathrm{C}} = \{1, 2, \cdots, L\} - \{t_1, t_2, \cdots, t_K\}$,
where ε_{t_i} is the ith smallest in set $\{\varepsilon_k\}$.

already been built in [13]. However, this work wants to emphasize a direct approach, which has been mentioned by Candès just as CS appeared [21]. Let's rewrite (1) in matrix formulation as

$$\begin{bmatrix} s_n \\ s_{n-1} \\ \vdots \\ s_{n-L+1} \end{bmatrix} = \begin{bmatrix} s_{n-1} & \cdots & s_{n-F} \\ s_{n-2} & \cdots & s_{n-F-1} \\ \vdots & \vdots & \vdots \\ s_{n-L} & \cdots & s_{n-L-F+1} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_F \end{bmatrix} + \begin{bmatrix} r_n \\ r_{n-1} \\ \vdots \\ r_{n-L+1} \end{bmatrix}, \quad (11)$$

or briefly

$$\mathbf{s} = \bar{\mathbf{S}}\bar{\mathbf{a}} + \mathbf{r}. \quad (12)$$

The Toeplitz matrix $\bar{\mathbf{S}}$ is generally of full rank and $\bar{\mathbf{S}}^{\mathrm{T}}$ has a nullspace with nullity $L - F$. One picks M $(M \leq L - F)$ orthogonal vectors $\{\mathbf{h}_i\}_{1 \leq i \leq M}$ in $\mathrm{null}(\bar{\mathbf{S}}^{\mathrm{T}})$ and constructs a matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_M]^{\mathrm{T}}$. By left multiplying $\mathbf{H}$ to the both sides of (12) and applying $\mathbf{H}\bar{\mathbf{S}} = \mathbf{0}$, one has

$$\mathbf{Hs} = \mathbf{Hr}. \quad (13)$$

By defining $\mathbf{y} = \mathbf{Hs}$, LPSR problem of (12) changes to

$$\hat{\mathbf{r}} = \arg\min \|\mathbf{r}\|_0, \quad \text{s.t.} \quad \mathbf{y} = \mathbf{Hr}, \quad (14)$$

which is a standard sparse recovery problem in CS. There are many algorithms to solve (14), two of which will be tested in next section. Considering the problem of whether $L - F$ measurements is enough to recover the desired non-zero residuals [22], one may try to compare the two approaches of solving a LPSR problem, or its equivalence in CS. However, this topic beyond the scope of this paper.

5. SIMULATION RESULTS

The proposed greedy algorithms (Greedy1 and Greedy2) are tested and compared with the reweighted l_1 algorithm (Reweight)[13] and two sparse recovery algorithms, OMP [23, 24] and ZAP [25]. Furthermore, the conventional LP of LS minimization is also conducted for comparison.

The test signal is a male speech in English with sampling rate 8KHz. The frame length and the all-pole filter order are 160 and 10, respectively. Totally 2500 frames including voiced frame and unvoiced frame are processed separately. For each frame, the support set of the residual vector is estimated by all algorithms, while the long term correlation in voiced frame is firstly removed by pitch estimation. To the reference methods, which do not output residual support, the residual is calculated and then the locations of the K largest entries are selected as $\hat{U}$. Then the prediction coefficient $\hat{\mathbf{a}}$ is estimated based on LS minimization on support $\hat{U}_{\mathrm{C}}$. The

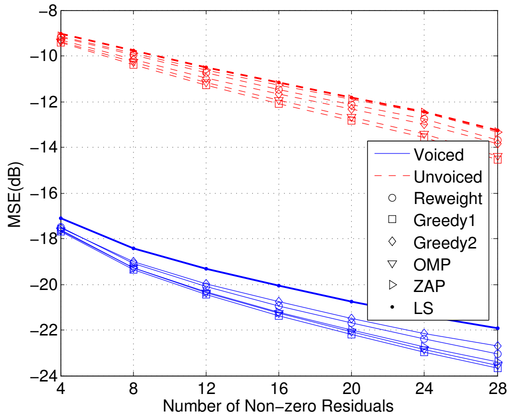

Fig. 2. MSEs in linear representation form v.s. non-zero residual numbers, where LS result is highlighted.

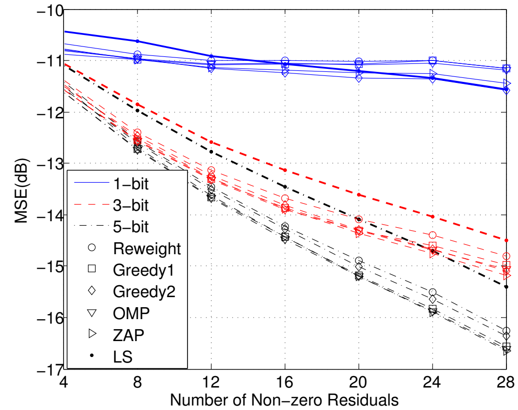

Fig. 3. MSEs of voiced synthetic speech with AbS v.s. non-zero residual numbers, where LS result is highlighted.

synthetic speech is the output of the estimated all-pole filter excited by K sparse residuals. To be noticed, the perceptually weighting is not considered for simplicity. Therefore, mean squared error between the synthetic signal and the original speech, $\mathrm{MSE} = 20\lg(\|\mathbf{x} - \hat{\mathbf{x}}\|_2/\|\mathbf{x}\|_2)$, is used for evaluation. In order to look into the problem and provide comprehensive comparison, several kinds of synthetic speeches are adopted in two experiments.

In the first experiment, the synthetic signal is generated by a linear representation formula of (11), where $\mathbf{a}$ and $\mathbf{r}$ are replaced by their estimates, while the historic speech vectors are from the original speech. Obviously this setting is unrealistic, because the original speech does not exist in the decoder side. However, this is the only approach that satisfies the optimization model of LPSR of (8) and (9). MSEs with respect to the predefined non-zero residual number are plotted in Fig. 2. There are two clusters of curves and MSEs of the voiced frames are much lower that those of the unvoiced frames, which verifies that voiced frame fits LPSR model better, while the unvoiced speech is much more unpredictable. For both clusters, MSEs decrease as non-zero residual number increases, which reveals that the LP residual of speech is not exactly sparse. In each cluster, MSE of conventional LP is the largest, as verified that sparse residual constraint could improve the accuracy of LP model. Furthermore, the proposed greedy methods, especially Greedy1, behaves pretty well and its MSE is the lowest in both voiced and unvoiced cases.

The second experiment is to provide more insight to LPSR problem in a more realistic scenario of speech coding. The non-zero residuals are estimated using Analysis by Synthesis (AbS) technique and then quantized before exciting the estimated all-pole filter. Each non-zero residual is quantized by $1, 3$, or 5 bits. MSEs of the voiced and unvoiced frames are plotted, respectively, in Fig. 3 and Fig. 4, with respect to different predefined non-zero residual number. Some observations can be obtained from both figures. First, the curves are clustered by numbers of quantization level, but 3-bit and 5-bit quantization are rather close. Second, in 3-bit and 5-bit quantization case of both voiced and unvoiced frames, conventional LS minimization results in the largest MSE among all solutions, especially for the voiced frames, and the proposed greedy algorithms behave pretty well. Furthermore, one can read that 5-bit quantization with LP is not as good as 3-bit quantization with LPSR, which proves that LPSR model provides better estimate and may save bits. Finally,

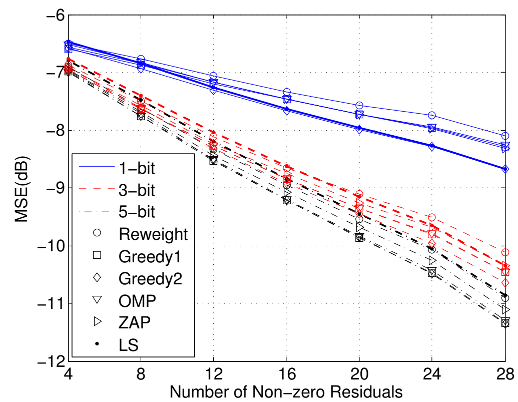

Fig. 4. MSEs of unvoiced synthetic speech with AbS v.s. non-zero residual numbers, where LS result is highlighted.

in 1-bit quantization case, however, it is hard to say whether LPSR model is better than the conventional LP. One should not be surprised by this result, considering that after such strong distortion in residual domain, the demand for an accurate all-pole filter is down. Combining the above observations, one may readily accept that LPSR model is better for moderate speech coder, in which more bits are provided for coding residuals.

6. CONCLUSION

In this paper, LPSR problem is reformulated as finding the largest linear correlated strict subset in a given vector set. A greedy approach is proposed to pick out those *outliers* iteratively based on a likelihood represented by temporary prediction error. The proposed methods are tested in the speech coding scenario and compared with conventional LS method, a recent proposed l_1 reweighted method, and two sparse signal recovery algorithms, by directly casting LPSR problem in the CS framework. Experiment results demonstrate that the proposed greedy methods work well. It is further shown clearly that LPSR model is more practical in the speech coders that more bits are allocated to the residuals.

7. REFERENCES

[1] B.S. Atal and M.R. Schroeder, "Predictive coding of speech," in *Proceedings of Conference on Communications and Processing*, 1967, pp. 360–361.

[2] S. Saito, Fukumura, and F. Itakura, "Theoretical consideration of the statistical optimum recognition of the spectral density of speech," *Journal of Acoustical Society of Japan*, January 1967.

[3] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*, John Wiley & Sons, Inc., 2003.

[4] B.S. Atal and M.R. Schroeder, "Stochastic coding of speech signals at very low bit rates," in *Proceedings of International Conference on Communications*, 1984, pp. 1610–1613.

[5] M.R. Schroeder and B.S. Atal, "Code-excited linear prediction (celp): High-quality speech at very low bit rates," in *ICASSP*, March 1985, pp. 937–940.

[6] E. Denoel and J.-P. Solvay, "Linear prediction of speech with a least absolute error criterion," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. ASSP-33, no. 6, pp. 1397–1403, 1985.

[7] H.-T. Hu, "Linear prediction using l_1 norm in orthogonal vector space," *Electronics Letters*, vol. 31, no. 6, pp. 430–431, March 1995.

[8] R. J. Mammone, K. Wang, and S. Gay, "Lpc speech analysis using the l_1 norm," in *ICASSP*. IEEE, 1985, vol. 10, pp. 485–488, IEEE.

[9] J. Lansford and R. Yarlagadda, "Adaptive l_p approach to speech coding," in *ICASSP*, New York, NY, April 1988, IEEE, vol. 1, p. 335, IEEE.

[10] C.-H. Lee, "On robust linear prediction of speech," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 36, no. 5, pp. 642–650, 5 1988.

[11] M. N. Murthi and B. D. Rao, "Towards a synergistic multistage speech coder," in *ICASSP*, Seattle, WA, May 1998, IEEE, vol. 1, pp. 369– 372, IEEE.

[12] Y. Li, "A globally convergent method for zp problems," *SIAM J. Optimization*, vol. 3, no. 3, pp. 609–629, August 1993.

[13] DD. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Transactions on Acoustic, Speech, and Language Processing*, vol. 20, no. 5, pp. 1644–1657, July 2012.

[14] E. J. Candés and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[15] R. Yarlagadda, J. B. Bednar, and T. L. Watt, "Fast algorithms for l_p deconvolution," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. ASSP-33, no. 1, pp. 174–182, February 1985.

[16] M. Namba, H. Kamata, and Y. Ishida, "Pole-zero estimation of speech lbased on l_1 norm linear prediction," in *Proceedings of IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing*, Victoria, BC, May 1995, IEEE, pp. 567–570.

[17] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear predictors for speech processing," *Proceedings of Interspeech*, pp. 1353–1356, 2008.

[18] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Retrieving sparse patterns using a compressed sensing framework: Applications to speech coding based on sparse linear prediction," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 103–106, January 2010.

[19] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. ASSP-26, no. 3, pp. 197–210, June 1978.

[20] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 3rd edition, 1996.

[21] E. J. Candés and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[22] D. Donoho and J. Tanner, "Counting faces of randomly-projected polytopes when the projection radically lowers dimension," *Journal of the American Mathematical Society*, pp. 1–53, 2009.

[23] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Conf. Rec. 27th Asilomar Conf. Sig., Sys., & Comput.*, November 1993, vol. 1, pp. 40–44.

[24] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Information Theory*, vol. 53, no. 12, pp. 4655–4666, December 2007.

[25] J. Jin, Y. Gu, and S. Mei, "A stochastic gradient approach on compressive sensing signal reconstruction based on adaptive filtering framework," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 409–420, April 2010.