A Multipulse Speech Coding Model via $\ell_{1/2}$–norm Minimization based Linear Prediction and Sparse Decomposition

Suiqing Xue [a*], Gang Min [b] and Guolei Ren [c]

College of Information and Communication, National University of Defense Technology, Wuhan, 430010 China

[a]175535882@qq.com, [b]mgxaty@gmail.com, [c]dzxxlab@163.com

Abstract. Sparse linear predictive analysis using the ℓ_1–norm minimization criterion has been shown to provide a valid alternative to the traditional linear predictive approach. The sparser the resulted speech residual is, the fewer pulses needed to represent it. To find a sparse residual further, we propose to use the $\ell_{1/2}$–norm as the optimization objective of linear prediction, and use an iteratively reweighted ℓ_1 minimization approach to solve it. We also find that the procedure of determining the locations and amplitudes of the optimal pulses in the multipulse analysis is equivalent to a sparse decomposition problem, which is efficiently solved by the optimized orthogonal matching pursuit approach. The objective and informal subjective evaluations over the TIMIT database give proof of the effectiveness of this new model, performing better than the traditional approach for modeling and coding of speech.

Keywords: Multipulse speech coding, $\ell_{1/2}$–norm, Sparse linear prediction, Sparse decomposition.

Introduction

Linear predictive (LP) analysis of speech is one of the most important speech analysis techniques, which is widely used in low bit rate speech coding, automatic speaker and speech recognition, etc[1][2]. LP analysis is closely related to the source-filter production model of speech, where a sampled speech signal can be modeled as the output of a linear, time-varying, all-pole filter excited by a excitation signal. Conventionally, the LP coefficients are identified via the ℓ_2–norm criteria, i.e., to minimize the energy of the LP residual of speech. The choice of ℓ_2–norm is popular because the closed-form optimal solution exists and it can be efficiently solved by the recursion methods, such as Levinson-Durbin method [3]. However, there are some obvious defects to use the ℓ_2–norm as the optimization objective. Firstly, the ℓ_2–norm resulted LP coefficients are highly sensitive to the outlier noises [4]. Moreover, this traditional LP model is not accurate for voiced speech where the excitation signal is sparse, quasi-periodic, and pulse-like [5].

In recent years, sparse LP analysis using the ℓ_1–norm criterion has attracted growing attention, which shows to provide a valid alternative to the ℓ_2–norm based linear prediction method. Sparse linear prediction provides a sparser residual. In addition, it can also effectively decouples the vocal tract filer from its excitation. These superior properties are beneficial to speech coding [6]. Except for the ℓ_1–norm based methods, the authors in [7] propose another efficient sparse LP analysis method. Although the LP residual is very sparse, an extra Glottal closure instants (GCI) detecting procedure must be performed at the beginning. The authors in [8] incorporate the sparse LP analysis into the well-known multipusle excitation (MPE) and algebraic code excited linear prediction (ACELP) speech coding model and achieve promising results.

The ℓ_0–norm is the best metric for sparsity, however, ℓ_0–norm optimization is a NP-hard problem. As a result, we use the $\ell_{1/2}$–norm as a valid alternative, which is supposed to provide sparser LP residual than the ℓ_2–norm and ℓ_1–norm. MPE is a powerful model for speech coding, which is the fundamental model for many speech coding standards, such as G.723, AMR, etc[9]. Whereas, the quality of the synthesized speech using the MPE model degrades dramatically if there are not enough pulses to describe the LP residual. On the one hand, the LP residual via the traditional ℓ_2–norm based linear prediction approach is not sparse. On the other hand, the multipulse analysis procedure also needs to be improved. Consequently, we propose to use the sparse decomposition approach to solve the challenging problem of determining the locations and amplitudes of the optimal pulses. These improvements mentioned above are natural extension of the existed techniques.

This paper is organized as follows. In section 2, we provide a prologue that describes the problem formulation of sparse linear prediction. In section 3, we propose the sparse decomposition based multipulse speech coding model. Experimental results are outlined in section 4. Section 5 concludes our work.

Sparse linear prediction via $\ell_{1/2}$–norm minimization

Linear predictive analysis of speech. The main idea of LP analysis is that the current sample of speech signals s(n) can be predicted by the P past samples,

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

where $\{a_k, k = 1, 2, ..., P\}$ are the P order prediction coefficients, $r(n)$ is the prediction error signals. $r(n)$ is also called the LP residual, which corresponds to the excitation in the autoregressive production model of speech.

Speech is usually processed frame by frame, as a result, we can write the LP analysis for a speech frame of N samples, i.e., $s(n)$, $n = 1, 2, ..., N$, in the following matrix form,

$$s = Sa + r \tag{2}$$

Then, the LP coefficients can be computed by solving the following (L_q) optimization problem,

$$(L_q) \qquad \hat{a} = \arg\min_a \|s - Sa\|_q^q \tag{3}$$

Here,

$$s = \begin{bmatrix} s(n_1) \\ \vdots \\ s(n_2) \end{bmatrix}, S = \begin{bmatrix} s(n_1 - 1) & \cdots & s(n_1 - P) \\ \vdots & \ddots & \vdots \\ s(n_2 - 1) & \cdots & s(n_2 - P) \end{bmatrix}$$
$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_P \end{bmatrix}, r = \begin{bmatrix} r(n_1) \\ \vdots \\ r(n_2) \end{bmatrix} \tag{4}$$

and the q–order norm $\|\cdot\|_q$ of $\boldsymbol{r}$ is defined as,

$$\|\boldsymbol{r}\|_q = (\sum_{n=1}^{N+P} |r(n)|^q)^{1/q} \tag{5}$$

where, n_1, n_2 denotes the starting and ending point of LP analysis, respectively. n_1, n_2 can be chosen in various ways by supposing $s(n) = 0$, $(n < 1, n > N)$. The autocorrelation method for LP analysis is a special case of (3), in which $q = 2$, $n_1 = 1$, $n_2 = N + P$. Given other choices of q, the estimated a_k and LP residual would have different properties.

Find a sparse residual via $\ell_{1/2}$–norm minimization. Under the ideal conditions, the sparsest residual can be obtained by minimize the ℓ_0–norm of the prediction error signals,

$$(L_0) \qquad \hat{a} = \arg\min_a \|s - Sa\|_0 \tag{6}$$

however, the (L_0) problem is highly non-convex, it is usually relaxed to a ℓ_1–norm minimization problem.

Furthermore, it is supposed to achieve much sparser residual by minimizing the ℓ_q–norm of the residual when $0 < q < 1$. Motivated by the fact that the $L_{1=2}$ regularizer has sparsity, unbiasedness and oracle properties [10], it seems very interesting for us to study the special case of $q = 1/2$ for linear prediction. Consequently, we obtain the following optimization problem,

$$(L_{1/2}) \qquad \hat{a} = \arg\min_a \|s - Sa\|_{1/2} \tag{7}$$

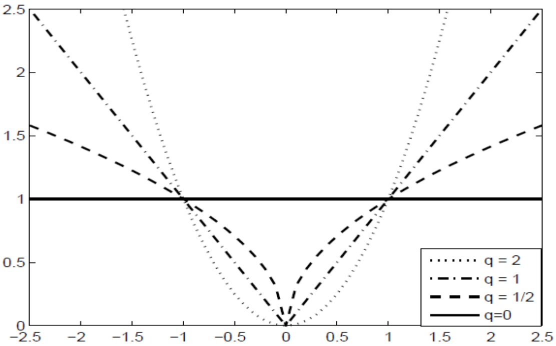

Fig. 1: Illustration of the cost function for q ≤ 2.

The $\ell_{1/2}$–norm based LP analysis for speech provides sparser LP residual when compared to the ℓ_2–norm and ℓ_1–norm based LP analysis, which can be directly explained in Fig. 1. With the increasing of q from 0 to 2, we can see that the penalty on outliers is also increasing. On the contrary, the penalty of the cost function on non-zero values approaches the ℓ_0–norm cost function, when $q \rightarrow 0$. Especially, the ℓ_0–norm weights all of the non-zero coefficients equally and provides the sparsest LP residual. However, the ℓ_0–norm minimization problem is highly non-convex and difficult to solve. As a result, we use the $\ell_{1/2}$–norm as an ideal approximation to ℓ_0–norm.

An approach for $\ell_{1/2}$–norm minimization. In this section, we propose to use a heuristic approach to solve the ($L_{1/2}$) problem, which is based on the minimization of an iteratively reweighted ℓ_1–norm measure [11], as is in (8).

$$\begin{aligned} \mathrm{a}^{(\mathrm{k}+1)} &= \arg\min_{\mathrm{a}} \|\mathrm{s} - \mathrm{Sa}\|_{1/2} = \arg\min_{a} \sum_{i=1}^{N} \sqrt{|s_i - (Sa)_i|} \\ &\approx \arg\min_{\mathrm{a}} \sum_{i=1}^{N} \frac{|s_i - (Sa)_i|}{\sqrt{|s_i - (Sa^k)_i|}} = \arg\min_{a} \|W^k(\mathrm{s} - \mathrm{Sa})\|_1 \end{aligned} \tag{8}$$

Firstly, we select an initial weight w_i, $i = 1, 2, \ldots, N$, and solve the weighted ℓ_1–norm minimization problem; Secondly, by using the solution, we update the weight matrix $\boldsymbol{W}$, which is a diagonal matrix and the diagonal element is w_i, then we solve the corresponding weighted ℓ_1–norm minimization problem again. We repeat these two steps until the iteration number k reaches the maximum iteration step K. The detailed description of solving the ($L_{1/2}$) problem is shown in algorithm1.

Multipulse speech coding via sparse decomposition of LP residual

In this section, we will incorporate the $\ell_{1/2}$–norm minimization based sparse LP analysis into the multipulse speech coding paradigm. The sparse LP residual can be represented via a signal with a limited number of pulses, i.e., only a few locations have the non-zero pulses. Speech can be synthesized with any desired quality by providing a sufficient number of pulses at the input of the all-pole synthetic filter $H(z)$ [9]. The remained problems are how many pulses and what the locations and amplitudes of these pulses should be. These questions are the core of the multipulse analysis.

Algorithm 1 Solving the ($L_{1/2}$) problem via iteratively reweighted ℓ_1 minimization.

```
1: Input: $S, s$
2: Output: $\hat{a}$
3: Initialization: $a^{(0)} = 0, w_i^{(0)} = 1, i = 1,2,\ldots,N, W^{(0)} = diag([w_1^{(0)}, w_2^{(0)}, \ldots, w_N^{(0)}]), k = 0, K$.
4: while $k < K$ do
5: // Line 6 solves the weighted $\ell_1$ minimization problem:
6: $a^{(k+1)} = arg\,min_a \|W^k(s - Sa)\|_1$
7: // Lines 8–9 update the weight matrix:
8: $w_i^{(k+1)} = 1/\sqrt{|s_i - (Sa^{k+1})_i| + \varepsilon}, i = 1,2,\ldots,N$
9: $\boldsymbol{W}(k+1) = diag([w_1^{(k+1)}, w_2^{(k+1)}, \ldots, w_N^{(k+1)}])$
10: $k = k + 1$
11: end while
12: $\hat{a} = a^{(k)}$
```

Multipulse coding of speech. In the traditional multipulse speech coding system, the locations and amplitudes of the optimum pulses are coded in the encoder and transmitted to the decoder. The determination of the optimal pulses is a combination optimization problem, which is NP-hard, so a suboptimal procedure is performed by the greedy algorithm. In practice, only one pulse is determined at a time, and then repeats the same procedure again. Fig. 2 illustrates the procedure for determining the optimum impulse locations. At the beginning, every location has the chance to be placed with the pulse in the current speech frame and this pulse is processed by *H*(*z*). The peak location is selected as the *k*th optimum pulse location. As for the corresponding optimum weighting coefficients, i.e., the pulse amplitude, it is calculated as [9],

$$\beta_k^{opt} = \frac{\sum_{n=0}^{N-1} \hat{e}^{(k-1)}(n)\,\hat{y}^{(k)}(n)}{\sum_{n=0}^{N-1}\left(\hat{y}^{(k)}(n)\right)^2} \tag{9}$$

Here, $\hat{y}^{(k)}(n)$ is the output of the vocal tract filter *H*(*z*) when excited by a single pulse, $\hat{e}^{(k-1)}(n)$ is the speech residual when subtracting out the effect of the formal $k-1$ pulses from the speech waveform.

Sparse decomposition based Multipulse speech coding. From a general view, we can model the multipulse analysis procedure as a sparse decomposition problem. For the *k*th speech frame $\boldsymbol{s}_k$, it can be divided into two parts, $\boldsymbol{u}_k$ and $\boldsymbol{e}_k$, where $\boldsymbol{u}_k$ is the delayed output of *H*(*z*) excited by the multipulse input of the $(k-1)$th speech frame and $\boldsymbol{e}_k$ is the present output of *H*(*z*) excited by the multipulse input of the *k*th speech frame.

$$\boldsymbol{s}_k = \boldsymbol{e}_k + \boldsymbol{u}_k \tag{10}$$

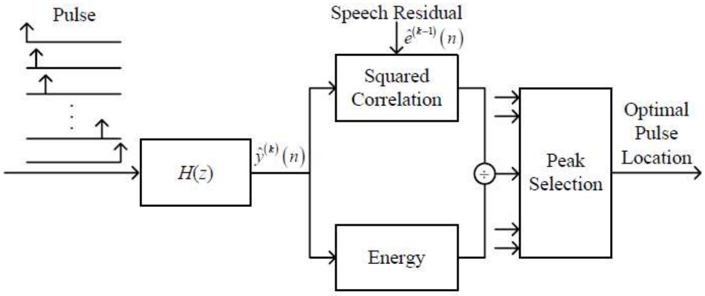

Fig. 2: Diagram illustrating the procedure for determining the optimum pulse locations in the multipulse analysis.

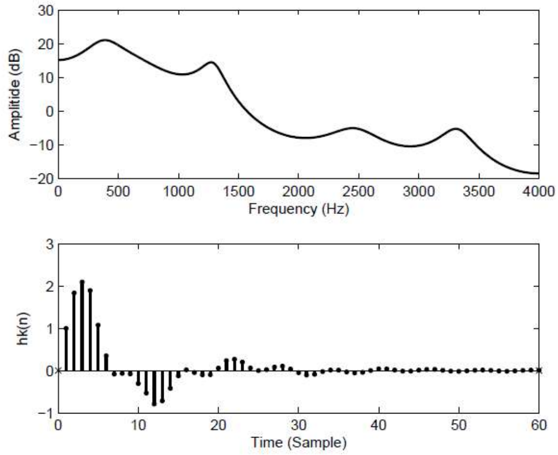

Fig. 3: Plots of the frequency response and pulse response of H(z) via $\ell_{1/2}$–norm linear predictor for a typical speech frame.

Let $h_k(n)$ denote the impulse response of *H(z)* for the *k*th speech frame. As is shown in Fig. 3, $h_k(n) \to 0$ with n↗. Thus, we suppose $h_k(n)$ is truncated to *N+1* coefficients. As the convolution between $h_k(n)$ and shifted pulses simply shifts $h_k(n)$ itself, we have,

$$\boldsymbol{u}_k = \boldsymbol{G}\boldsymbol{x}_{k-1} \tag{11}$$

$$\boldsymbol{e}_k = \boldsymbol{Hx}_k \tag{12}$$

where x_k is the multipulse excitation of the k^{th} speech frame, which contains only a few non-zero pulses, $\boldsymbol{G},\boldsymbol{H} \in \boldsymbol{R}^{N\times N}$,

$$G = \begin{bmatrix} h_{k-1}(N+1) & h_{k-1}N & \Lambda & h_{k-1}(2) \\ 0 & h_{k-1}(N+1) & \Lambda & h_{k-1}(3) \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & 0 & \Lambda & h_{k-1}(N+1) \end{bmatrix} \qquad H = \begin{bmatrix} h_k(1) & 0 & \Lambda & 0 \\ h_k(2) & h_k(1) & \Lambda & 0 \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ h_k(N) & h_k(N-1) & 0 & h_k(1) \end{bmatrix}$$

For the k^{th} speech frame, s_k is known while $\boldsymbol{u}_k$ can be computed from (11). Our goal is to use the least pulses and achieve the best speech quality. Consequently, we have the following sparse decom-position problem to determine $\boldsymbol{x}_k$,

$$\begin{gathered} \min \ \|\boldsymbol{x}_k\|_0 \\ \text{subject to } \boldsymbol{s}_k - \boldsymbol{u}_k = \boldsymbol{Hx}_k \end{gathered} \tag{13}$$

In a word, the application of sparse decomposition to model the LP residual is shown in **algorithm2**.Here, only step 16 remains unresolved, we will detail it in the next section.

Implementation of sparse decomposition via OOMP. Step 16 in **algorithm2** is a standard sparse decomposition problem, let us rewrite it briefly. If $\mathcal{H}$ denotes a Hilbert space, $\boldsymbol{\Gamma}$ denotes a set of indices, then the well-known sparse decomposition problem can be described as follows,

$$\begin{gathered} \min \ \|\boldsymbol{x}\|_0 \\ \text{subject to } \boldsymbol{f} = \boldsymbol{\mathcal{D}x} \end{gathered} \tag{14}$$

where $\boldsymbol{f}$, $\boldsymbol{\mathcal{D}} = \{\boldsymbol{\alpha}_n,\ n \in \Gamma\ \}$ denotes the signal and dictionary, respectively, $\boldsymbol{x}$ is the sparse representation vector, $\boldsymbol{\alpha}_n$ n is the atom in the dictionary. Conventionally, $\boldsymbol{\mathcal{D}}$ is a redundant dictionary, i.e., there are more atoms than the dimension of the decomposed signal.

There are many approaches on the shelf to solve the sparse decomposition problem shown in (14), such as matching pursuit (MP), orthogonal matching pursuit (OMP), non-convex regularization, etc. Here, we introduce a natural improvement to the MP approach, called optimized orthogonal matching pursuit (OOMP) [12], to solve it.

Algorithm 2 Multipulse speech coding via sparse decomposition of LP residual.

```
1: Input: speech signals, s(n)
2: Output: multipulse vector, $\boldsymbol{x}_k$, $k = 1,2,...,M$
3: Initialization: $\boldsymbol{x}_0=0$, h0(n)=0, n=1,2,...,$N$, $k=1$, number of speech frames $M$, window shift R.
4: for k = 1 to $M$ do
5: // Line 6 enframes speech $s(n)$ by a window $w(n)$:
6: $s_k(n) = s(kR + n)w(n)$
7: // Line 8 performs the sparse linear predictive analysis:
8: $\hat{a}_k = \arg\min_a \|\boldsymbol{s}_k - \boldsymbol{S}_k\boldsymbol{a}\|_{1/2}$
9: // Lines 10–11 compute the impulse response of the all-pole synthetic filter of the k th frame:
10: $H_k(z) = 1/\left(1 - \sum_{i=1}^{P} \hat{a}_{ki} z^{-i}\right)$
11: $h_k(n)=Z^{1}(Hk(z))$
12: // Line 13 computes the delayed output excited by the multipulse input of the $(k-1)^{th}$ frame:
13: $\boldsymbol{u}_k = \boldsymbol{Gx}_{k-1}$
14: // Line 15 computes the objective synthetic speech:
15: $\boldsymbol{e}_k = \boldsymbol{s}_k - \boldsymbol{u}_k$
16: Get $\boldsymbol{x}_k$ by solving the problem in (13)
17: $k = k + 1$
18: end for
```

Experimental results

In the next experiments, we randomly select 90 speech utterances spoken by 15 males and 15 females from the TIMIT database. Each utterance is about 3 seconds in duration, which is down-sampled at 8 kHz. The speech frame is 5ms *(N* = 40). A 10 order (*P* = 10) linear predictor is performed to each frame with 20 samples looking ahead and 20 samples looking back, i.e., the length of the hamming window for LP analysis is 80 samples, while the window shift is 40 samples.

PESQ is a popular objective measure of speech quality, which is developed to obtain the highest correlation with subjective MOS and standardized by ITU as P.862 [13]. The frequency-weighted segmental SNR (fwsegSNRs) is another widely used objective measure for speech processing. For both of these measures, higher value indicates better performance. In the tests, we will utilize both of these measures to evaluate the quality of the synthesized speech.

Linear predictive residual of speech. The (L 1/2) problem is solved by **algorithm1** where the param-eters are as follows, $\varepsilon = 10^{-5}$, 20 iterations (K = 20) is empirically adequate to achieve a stationary solution. The ℓ_1–norm minimization is implemented by the ℓ_1–magic toolbox. Fig. 4 illustrates the LP residual for different LP analysis approaches. It can be seen that the LP residual predicted by the $\ell_{1/2}$–norm based LP approach is much sparser, and it is elongated and more identical to a multipuse signal.

Speech quality evaluation. In this section, we will evaluate the quality of the synthesized speech via the fwsegSNRs and PESQ measures. The traditional MPE model and the recently reported ℓ_1–norm based multipulse coding models are involved for comparison. The residual model in these schemes is briefly described in Tab.1.

In Tab.2, $\mathbf{LP}_2$ denotes the traditional ℓ_2 –norm based linear prediction, $\mathbf{SLP}_1$, $\mathbf{SLP}_{1/2}$ denotes the ℓ_1–norm and $\ell_{1/2}$–norm based sparse linear prediction, respectively. The number of pulses in the multipulse analysis is 5 and 10, i.e., T = 5,10. We focus on the whole speech coding paradigm, so the locations and amplitudes of the optimal pulses, LP coefficients are all left unquantized. We can see that the proposed speech coding model via $\mathbf{SLP}_{1/2}$ and OOMP substantially achieves better performance than other models, where the fwsegSNRs is improved by 2.4dB and 4dB for T = 5,10, respectively. Also, the PESQ score is at least improved by 0.26 and 0.3. The informal subject listening tests also demonstrate that the synthesized speech using the proposed model is more clear and intelligible.

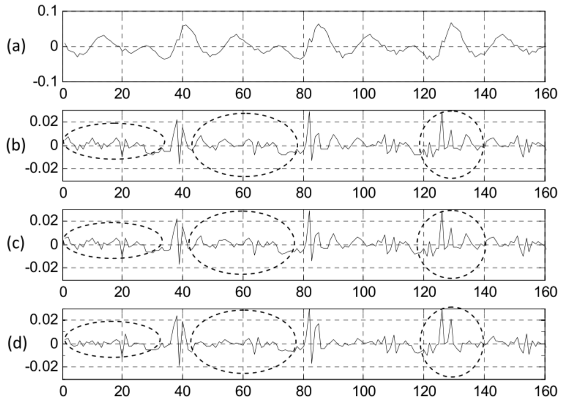

Fig. 4: Plots of the LP residual via different linear predictors. (a) speech; (b) ℓ_2–norm; (c) ℓ_1–norm;(d) $\ell_{1/2}$–norm.

Table 1: Description of the different residual models compared in the evaluation.

Residual model	Description
MPE	Multipulse excitation model [9]
MaxSa	T pulses of largest magnitude in residual
AoMaxSa	Amplitudes optimized **MaxSa** [5][8]
OOMP	Optimized orthogonal matching pursuit

Summary

In this paper, we propose a new speech coding model via $\ell_{1/2}$–norm minimization based LP analysis and sparse decomposition. To find a sparser LP residual, we use the $\ell_{1/2}$–norm as the optimization objective, and take the iteratively reweigthed ℓ_1 minimization approach to solve it. The multipulse analysis is implemented by the OOMP approach. The experimental results over the TIMIT database are quite promising. We will quantize the sparse LP coefficients, locations and amplitudes of the optimal pulses to fix the speech coding rate in future work.

Table 2: Comparison on the quality of the synthesized speech using different speech coding models.

LP method	Residual model	*T*	fwsegSNRs	PESQ
LP_2	**MPE**	5	14.282	3.322
SLP_1	**MaxSa**	5	14.265	3.116
SLP_1	**AoMaxSa**	5	14.277	3.235
$SLP_{1/2}$	**OOMP**	5	**16.676**	**3.581**
LP_2	**MPE**	10	14.768	3.536
SLP_1	**MaxSa**	10	17.517	3.454
SLP_1	**AoMaxSa**	10	18.418	3.659
$SLP_{1/2}$	**OOMP**	10	**22.515**	**3.993**

References

[1] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” Journal of the Acoustical Socielty of America, vol. 50, pp. 637–655, 1971.

[2] J. Makhoul, “Linear prediction: a tutorial review,” Proceedings of the IEEE, vol. 63, no. 4, pp. 561–580, 1975.

[3] B. S. Atal, “The history of linear prediction,” IEEE Signal Processing Magazine, vol. 23, no. 2, pp. 154–161, 2006.

[4] C. H. Lee, “On robust linear prediction of speech,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 36, no. 5, pp. 642–650, 1988.

[5] D. 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 Audio, Speech and Language Processing, vol. 20, no. 5, pp. 1644–1657, 2012.

[6] D. Giacobello, M. G. Christensen, T. L. Jensen, M. N. Murthi, s. H. Jensen, M. Moonen, “Stable 1-norm error minimization based linear predictors for speech modeling,” IEEE Transactions on Audio, Speech and Language Processing, vol. 22, no. 5, pp. 912–922, 2014.

[7]V.Khanagha and K.Daoudi,“An efficient solution to sparse linear prediction analysis of speech,” EURASIP Journal on Audio, Speech and Music Processing, vol. 3, pp. 1–9, 2013.

[8] G. Alipoor and M. H. Savoji, “Wide-band speech coding based on bandwidth extension and sparse linear prediction,” in TSP 2012 – 35th International Conference on Telecommunications and Signal Processing, July 03–04, Prague, Czech Republic, Proceedings, 2012, pp. 454–459.

[9] B. S. Atal and J. R. Remde, “A new model of LPC excitation for producing natural sounding speech at low bit rates,” in ICASSP 1982 – 7th International Conference on Acoustics, Speech, and Signal Processing, May, USA, Proceedings, 1982, pp. 614–617.

[10] Z. B. Xu, Z. Hai, W. Yao, C. X. Yu, and L. Yong, “$L_{1/2}$ regularization,” Science China (Information Sciences), vol. 53, no. 6, pp. 1159–1169, 2010.

[11] E. J. Candes, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted ℓ_1 minimizaton,” Journal of Fourier Analysis and Application, vol. 14, pp. 877–905, 2008.

[12] L.Rebollo-Neira and D.Lowe, “Optimized or thogonal matching pursuit appraoch,” IEEE Signal Processing Letters, vol. 9, no. 4, pp. 137–140, 2002.