RE-ESTIMATION OF LINEAR PREDICTIVE PARAMETERS IN SPARSE LINEAR PREDICTION

Daniele Giacobello[1], *Manohar N. Murthi*[2], *Mads Græsbøll Christensen*[3], *Søren Holdt Jensen*[1], *Marc Moonen*[4]

[1]Dept. of Electronic Systems, Aalborg Universitet, Aalborg, Denmark
[2]Dept. of Electrical and Computer Engineering, University of Miami, USA
[3]Dept. of Media Technology, Aalborg Universitet, Aalborg, Denmark
[4]Dept. of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium
{dg,shj}@es.aau.dk, mmurthi@miami.edu, mgc@imm.aau.dk, marc.moonen@esat.kuleuven.be

ABSTRACT

In this work, we propose a novel scheme to re-estimate the linear predictive parameters in sparse speech coding. The idea is to estimate the optimal truncated impulse response that creates the given sparse coded residual without distortion. An all-pole approximation of this impulse response is then found using a least square approximation. The all-pole approximation is a stable linear predictor that allows a more efficient reconstruction of the segment of speech. The effectiveness of the algorithm is proved in the experimental analysis.

1. INTRODUCTION

The most important speech coding paradigm in the past twenty years has been *Analysis-by-Synthesis* (AbS) [1, 2]. The name signifies analysis of the optimal parameters by synthesizing speech based on these. In other words, the speech encoder mimes the behavior of the speech decoder in order to find the best parameters needed. The usual approach is to first find the linear prediction parameters in a open-loop configuration then searching for the best excitation given certain constraints on it. This is done in a closed-loop configuration where the perceptually weighted distortion between the original and synthesized speech waveform is minimized. The conceptual difference between a quasi-white true residual and its approximated version, where usually sparsity is taken into consideration, creates a mismatch that can raise the distortion significantly. In our previous work we have defined a new synergistic predictive framework that reduces this mismatch by jointly finding a sparse prediction residual as well as a sparse high order linear predictor for a given speech frame [3]. Multipulse encoding techniques [4] have shown to be more consistent with this kind of predictive framework, offering a lower distortion with very few samples [5].

In this work, we propose a method to further reduce the mismatch between sparse linear predictor and approximated residual by re-estimating the linear predictive parameters. This paper is structured as follow. In Section 2, we introduce the coding method based on sparse linear prediction. In Section 3, we introduce the re-estimation procedure and in Section 4 we propose the results to validate our method. Finally, Section 5 concludes our work.

2. SPEECH CODING BASED ON SPARSE LINEAR PREDICTION

In our previous work [3, 5], we have defined a synergistic new predictive framework that jointly finds a sparse prediction residual $\mathbf{r}$ as well as a sparse high order linear predictor $\mathbf{a}$ for a given speech frame $\mathbf{x}$ as

$$\hat{\mathbf{a}}, \hat{\mathbf{r}} = \arg\min_{\mathbf{a}} \|\mathbf{r}\|_1 + \gamma\|\mathbf{a}\|_1, \quad \text{subject to} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}; \tag{1}$$

where:

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

and $\|\cdot\|_1$ is the 1-norm defined as the sum of absolute values of the vector on which operates. The start and end points N_1 and N_2 can be chosen in various ways assuming that $x(n) = 0$ for $n < 1$ and $n > N$ [6]. The more tractable 1-norm $\|\cdot\|_1$ is used as a linear programming relaxation of the sparsity measure, often represented as the cardinality of a vector, the so-called 0-norm $\|\cdot\|_0$. This optimization problem can be posed as a linear programming problem and can be solved using an interior-point algorithm [7]. The choice of the regularization term γ is given by the L-curve where a

The work of Daniele Giacobello is supported by the Marie Curie EST-SIGNAL Fellowship (http://est-signal.i3s.unice.fr), contract no. MEST-CT-2005-021175.

The work of Manohar N. Murthi is supported by the National Science Foundation via awards CCF-0347229 and CNS-0519933.

trade-off between the sparsity of the residual and the sparsity of the predictor is found [8].

The sparse structure of the predictor allows a joint estimation of short-term and long-term predictor [9]:

$$A(z) \approx \tilde{A}(z) = F(z)P(z), \tag{2}$$

where $F(z)$ is the short-term predictor, commonly employed to remove short-term redundancies due to the formants, and $P(z)$ is the pitch predictor that removes the long-term redundancies. The sparse structure of the true residual $\hat{\mathbf{r}}$ allows for a quick and more efficient search of approximated residual $\tilde{\mathbf{r}}$ using sparse encoding procedure, where the approximated residual is given by a regular pulse excitation (RPE) [10]. The problem can be rewritten as:

$$\tilde{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{W}(\mathbf{x} - \tilde{\mathbf{H}}\mathbf{r})\|_2, \tag{3}$$

by imposing the RPE structure on $\tilde{\mathbf{r}}$:

$$\tilde{r}(n) = \sum_{i=0}^{N/S-1} \alpha_i \delta(n - iS - s) \quad s = 0, 1, \ldots, S-1, \tag{4}$$

where α_i are the amplitudes $\delta(\cdot)$ is the Kronecker delta function, N/S are the number of pulses and S is the spacing; only S different configurations of the positions are allowed (s is the shift of the residual vector grid). In (3), $\mathbf{W}$ is the perceptual weighting matrix, $\tilde{\mathbf{H}}$ is the $(N) \times (K+N)$ synthesis matrix whose $i-$th row contains the elements of index $[0, K+i-1]$ of the truncated impulse response $\tilde{\mathbf{h}}$ of the combined prediction filter $\tilde{A}(z) = F(z)P(z)$:

$$\tilde{\mathbf{H}} = \begin{bmatrix} \tilde{h}_K & \cdots & \tilde{h}_0 & 0 & 0 & \cdots & 0 \\ \tilde{h}_{K+1} & \ddots & \ddots & \ddots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \cdots & \tilde{h}_0 & 0 & 0 \\ \tilde{h}_{K+N-2} & \ddots & \ddots & \cdots & \tilde{h}_1 & \tilde{h}_0 & 0 \\ \tilde{h}_{K+N-1} & \tilde{h}_{K+N-2} & \cdots & \cdots & \tilde{h}_2 & \tilde{h}_1 & \tilde{h}_0 \end{bmatrix}. \tag{5}$$

and $\mathbf{r}$ is composed of the previous residual samples $\tilde{\mathbf{r}}_-$ (the filter memory, already quantized) and the current $\tilde{\mathbf{r}}$ that has to be estimated:

$$\mathbf{r} = \begin{bmatrix} \tilde{\mathbf{r}}_-^T & \tilde{\mathbf{r}}^T \end{bmatrix}^T = [\tilde{r}_{-K}, \cdots, \tilde{r}_{-2}, \tilde{r}_{-1}, \tilde{r}_0, \tilde{r}_1, \tilde{r}_2, \cdots, \tilde{r}_{N-1}]^T. \tag{6}$$

In the end a segment of speech can be represented by the sparse predictor $\tilde{A}(z)$ and its approximated excitation $\tilde{\mathbf{r}}$.

3. RE-ESTIMATION OF THE PREDICTIVE PARAMETERS

To ensure simplicity in the following derivations, let's assume that no perceptual weighting is performed ($\mathbf{W} = \mathbf{I}$). The results can then be generalized for an arbitrary $\mathbf{W}$. The problem in (3) is now just a waveform matching problem. The interesting thing is that, once found a proper sparse excitation, we can re-estimate the matrix $\mathbf{H}$ and therefore the impulse response $\mathbf{h}$ by posing it as a convex optimization problem:

$$\hat{\mathbf{H}} = \arg\min_{\mathbf{H}} \|(\mathbf{x} - \mathbf{H}\tilde{\mathbf{r}})\|_2 \rightarrow \hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|(\mathbf{x} - \tilde{\mathbf{R}}\mathbf{h})\|_2 \tag{7}$$

where:

$$\tilde{\mathbf{R}} = \begin{bmatrix} \tilde{r}_0 & \cdots & \tilde{r}_{-K} & 0 & 0 & \cdots & 0 \\ \tilde{r}_1 & \ddots & \ddots & \ddots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \cdots & \ddots & 0 & 0 \\ \tilde{r}_{N-1} & \ddots & \ddots & \cdots & \ddots & \tilde{r}_{-K} & 0 \\ \tilde{r}_N & \tilde{r}_{N-1} & \cdots & \cdots & \cdots & \tilde{r}_{-K+1} & \tilde{r}_{-K} \end{bmatrix}. \tag{8}$$

where $\{\tilde{r}_{-K}, \ldots, \tilde{r}_{-1}\}$ is the past excitation (belonging to the previous frame). The problem (7) allows for a closed form solution when the 2-norm is employed in the minimization:

$$\hat{\mathbf{h}} = \mathbf{h}_{opt} = \tilde{\mathbf{R}}^T(\tilde{\mathbf{R}}\tilde{\mathbf{R}}^T)^{-1}\mathbf{x}. \tag{9}$$

Because the matrix $\tilde{\mathbf{R}}^T(\tilde{\mathbf{R}}\tilde{\mathbf{R}}^T)^{-1}$ in (9) is the pseudo-inverse $\tilde{\mathbf{R}}^+$ of $\tilde{\mathbf{R}}$, the new $\mathbf{h}_{opt}$ is then the optimal truncated impulse response that matches the given sparse residual:

$$\|\mathbf{x} - \tilde{\mathbf{R}}\mathbf{h}_{opt}\|_2 = 0. \tag{10}$$

It is therefore clear that the optimal sparse linear predictor $A(z)$ is the one that has $\mathbf{h}_{opt}$ as truncated impulse response. The problem now is that the impulse response will include both short-term and long-term contribution. We can split the two contribution and perform a two step optimization.

Assuming $\mathbf{h}_f$ the impulse response of the short-term predictor $1/F(z)$ and $\mathbf{h}_p$ the impulse response of the long-term predictor $1/P(z)$, we can rewrite the problem in (7) as:

$$\hat{\mathbf{H}}_f, \hat{\mathbf{H}}_p = \arg\min_{\mathbf{H}_f, \mathbf{H}_p} \|(\mathbf{x} - \mathbf{H}_f\mathbf{H}_p\tilde{\mathbf{r}})\|_2. \tag{11}$$

We can then proceed with the re-estimation of the impulse response of the short-term predictor by solving the problem:

$$\hat{\mathbf{h}}_f = \arg\min_{\mathbf{h}_f} \|(\mathbf{x} - (\mathbf{H}_p\tilde{\mathbf{R}})\mathbf{h}_f)\|_2, \tag{12}$$

and then find the predictor that approximates $\hat{\mathbf{h}}_f$. The predictor $A(z) = 1 + \sum_{k=1}^{Q} a_k z^{-k}$ can then just be seen as a reduced Q order IIR approximation ($Q << N + K$) to the optimal FIR filter $H_f(z)$. Assuming:

$$H_f(z) = \frac{E(z)}{A(z)} \tag{13}$$

where $E(z)$ is the error polynomial and $A(z)$ is the approximating polynomial:

$$E(z) = \sum_{k=0}^{N+Q-1} e_i z^{-i} \tag{14}$$

and

$$e_i = h_i^f - \sum_{k=1}^{Q} a_k h_{i-k}^f. \quad (15)$$

We recognize this also as a linear predictive problem. Putting (15) into matrix form:

$$\hat{\mathbf{e}} = \mathbf{h}_f - \mathbf{H}_f^F \hat{\mathbf{a}}, \quad (16)$$

and:

$$\mathbf{h}_f = \begin{bmatrix} h_f(N_1) \\ \vdots \\ h_f(N_2) \end{bmatrix}, \mathbf{H}_f^F = \begin{bmatrix} h_f(N_1-1) & \cdots & h_f(N_1-Q) \\ \vdots & & \vdots \\ h_f(N_2-1) & \cdots & h_f(N_2-Q) \end{bmatrix}$$

we can solve it using common procedures. In particular, rewriting the problem as:

$$\hat{\mathbf{a}} = \arg\min_{\hat{\mathbf{a}}} \|\mathbf{h}_f - \mathbf{H}_f^F \hat{\mathbf{a}}\|_2. \quad (17)$$

Choosing $N_1 = 1$ and $N_2 = N+Q$ and assuming $h_f(n) = 0$ for $n < 1$ and $n > N$, we find the well known Yule-Walker equations. This guarantees stability and simplicity of the solution. In more general terms the problem of approximating the impulse response $H_f(z)$ through the linear predictor $A(z)$ falls in the class of the approximation of FIR through IIR digital filters (see, for example, [12, 13]). Using a similar approach we can recalculate the long-term predictor as well.

4. EXPERIMENTAL ANALYSIS

In order to evaluate our method, we have analyzed about one hour of clean speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, re-sampled at 8 kHz. We choose a frame length of $N = 160$ (20 ms) and a order of the optimization problem in (1) of $K = 110$. We implement the sparse linear predictive coding using $N_f = 10$ and $N_p = 1$, the residual is encoded using RPE with 20 samples (pulse spacing $S = 8$), a gain and a shift. The gain is coded with 6 bits and the pulse amplitude are coded using a 8 level uniform quantizer, the LSF vector is encoded with 20 bits (providing transparent coding) using the procedure in [14], the pitch period is coded with 7 bits and the gain with 6 bits. This produces a fixed rate of 102 bit/frame (5100 bit/s). No perceptual weighting is employed. The re-estimation is done only on the short-term parameters. The coder that employs re-estimation consists of the following steps:

1. Determine $\tilde{A}(z) = F(z)P(z)$ using sparse linear prediction.
2. Calculate the residual vector $\tilde{\mathbf{r}}$ using RPE encoding.
3. Re-estimate the optimal truncated impulse response $\mathbf{h}_f$.

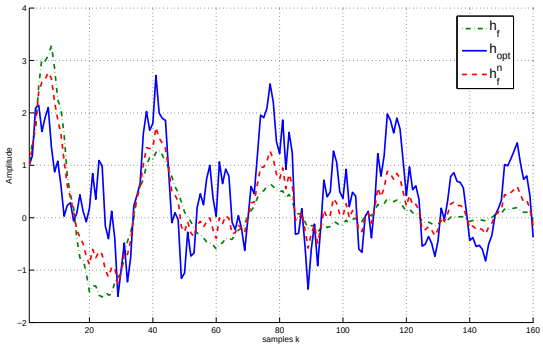

Fig. 1. *An example of the different impulse response used in the work. The impulse response* $\mathbf{h}_f$ *of the original short-term predictor* $F(z)$*, the optimal re-estimated impulse response adapted to the quantized residual* $\mathbf{h}_{opt}$ *and the approximated impulse response* $\mathbf{h}_f^n$ *of the new short-term predictor* $\hat{F}(z)$*. The order is* $N_f = 10$*.*

4. Least square IIR approximation of $\mathbf{h}_f$ using order $N_f = 8, 10, 12$.
5. Optimize the amplitudes of the sparse RPE residual $\tilde{\mathbf{r}}$ using the new synthesis filter $\hat{\mathbf{h}}_f$ (positions and shift stay the same).

We compare two approaches, one with only the re-estimation of $\mathbf{h}_f$ and one with the optimization of the amplitudes of the RPE residual, using (3). The results, in comparison with standard Sparse Linear Prediction, are shown in table 1. An example of the re-estimated impulses responses are shown in Figure 1.

Table 1. Improvements over conventional SPARSE LP in the decoded speech signal in terms of reduction of log magnitude segmental distortion (ΔDIST) and Mean Opinion Score (ΔMOS) using PESQ evaluation. A 95% confidence intervals is given for each value.

METHOD	ΔDIST	ΔMOS
N_f=8	+0.12$\pm$0.02 dB	+0.01$\pm$0.00
N_f=10	+0.35$\pm$0.03 dB	+0.05$\pm$0.00
N_f=12	+0.65$\pm$0.02 dB	+0.04$\pm$0.00
N_f=8 + REST	+0.17$\pm$0.01 dB	+0.03$\pm$0.00
N_f=10 + REST	+0.41$\pm$0.02 dB	+0.06$\pm$0.00
N_f=12 + REST	+0.71$\pm$0.04 dB	+0.07$\pm$0.00

5. CONCLUSIONS

In this paper, we have proposed a new method for the re-estimation of the prediction parameters in speech coding. In particular, the autoregressive modeling is no more employed as a method to remove the redundancies of the speech segment but as IIR approximation of the optimal FIR filter, adapted to the quantized approximated residual, that is used in the synthesis of the speech segment. The method has shown an improvement in the general performances of the sparse linear prediction framework, but it can be applied also to common methods based on minimum variance linear prediction (e.g. ACELP). The work can be extended for these methods where we expect an even greater increase in performances due to the mismatch between true residual and approximated one.

6. REFERENCES

[1] J. H. L. Hansen, J. G. Proakis, and J. R. Deller, Jr., *Discrete-time processing of speech signals*, Prentice-Hall, 1987.

[2] P. Kroon and W. B. Kleijn, "Linear-prediction based analysis-by-synthesis coding", in *Speech Coding and Synthesis*, Elsevier Science B.V., ch. 3, pp. 79–119, 1995.

[3] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse linear predictors for speech processing", *Proc. INTERSPEECH*, pp. 1353–1356, 2008.

[4] W. C. Chu, *Speech coding algorithms: foundation and evolution of standardized coders*, Wiley, 2003.

[5] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen "Speech Coding Based On Sparse Linear Prediction", to appear in *Proc. European Signal Processing Conference*, pp. 2524-2528, 2009.

[6] P. Stoica and R. Moses, *Spectral analysis of signals*, Pearson Prentice Hall, 2005.

[7] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[8] P. C. Hansen and D. P. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems", *SIAM Journal on Scientific Computing*, vol. 14, no. 6, pp. 1487–1503, 1993.

[9] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Joint estimation of short-term and long-term predictors in speech coders", in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 4109–4112, 2009.

[10] P. Kroon, E. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation - a novel approach to effective and efficient multipulse coding of speech", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1054–1063, 1986.

[11] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 221–239, 2000.

[12] H. Brandenstein, R. Unbehauen, "Least-squares approximation of FIR by IIR digital filters", *IEEE Trans. on Signal Processing*, vol. 46, pp. 21-30, 1998.

[13] B. Beliczynski, J. Kale, and G. D. Cain, "Approximation of FIR by IIR digital filters: An algorithm based on balanced model reduction", *IEEE Trans. Signal Processing*, vol. 40, pp. 532-542, 1999.

[14] A. D. Subramaniam, B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies", *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 2, pp. 130–142, 2003.