ESTIMATION OF FRAME INDEPENDENT AND ENHANCEMENT COMPONENTS FOR SPEECH COMMUNICATION OVER PACKET NETWORKS

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

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

ABSTRACT

In this paper, we describe a new approach to cope with packet loss in speech coders. The idea is to split the information present in each speech packet into two components, one to independently decode the given speech frame and one to enhance it by exploiting inter-frame dependencies. The scheme is based on sparse linear prediction and a redefinition of the analysis-by-synthesis process. We present Mean Opinion Scores for the presented coder with different degrees of packet loss and show that it performs similarly to frame dependent coders for low packet loss probability and similarly to frame independent coders for high packet loss probability. We also present ideas on how to make the coder work synergistically with the channel loss estimate.

Index Terms— Speech coding, Voice over IP (VoIP), linear predictive coding, analysis-by-synthesis.

1. INTRODUCTION

With the increasing importance of VoIP (Voice over IP) telephony, alternative methods to improve the robustness of speech codecs to packet loss are required. The approaches presented in literature, notably [1] with the definition of the iLBC (Internet Low Bit Rate Codec), tend to create speech coders that are totally frame independent or, in other words, where each frame is independently decodable and does not depend on the previous frames. On the other hand, in the case of telephony with dedicated circuits, the coding schemes used achieve high quality with low bit rate mostly because of their property to exploit inter-frame dependencies. However, these coding schemes, and in particular the ACELP (Algebraic Code Excited Linear Prediction) based codecs, in the case of packet loss show severe shortcomings [1].

In this paper we introduce a new approach to speech coding over packet networks, creating a coder that has frames with a core that is independently decodable and an enhancement layer that is based on the previously received frames. In particular, we create a coder that can select between two decoding procedures, if the previous frames are received correctly, then it decodes using all the information, otherwise, it uses only the frame independent information. By doing so, we offer the flexibility of a frame independent codec if the loss probability is significant but, if the probability is low (or ideally null), then it will exploit inter-frame dependencies to perform similarly to a frame dependent coder. In our coding scheme, the speech analysis is based on sparse linear prediction which has shown better statistical modeling in creating an output (residual and predictor) that offers better coding properties [2]. Frame independence is achieved through a rethinking of the analysis-by-synthesis (AbS) scheme [3], allowing the possibility of re-estimating the synthesis matrix (and thus the impulse response that generates it) that creates an independently decodable frame of speech given the residual similarly to what is done in [4].

The paper is organized as follows. Section 2 describes the system architecture of our coder. In Section 3, we provide some experimental results in comparison with G.729a [5] and iLBC, chosen due to their public availability. In Section 4, we discuss how the bit allocation can work synergistically with the channel loss statistics to generally improve the performance of the coder. Section 5 concludes our paper.

2. SYSTEM ARCHITECTURE

2.1. Step 1: Prediction parameters estimation

The first step is to perform a linear predictive analysis using a sparse linear prediction framework. A sparse linear predictive framework has already shown to offer, not only sparsity properties that make coding more straightforward [2] but also a more compact description of all the features extracted from a speech frame [7]. For a given speech frame $\mathbf{x}$, we obtain an estimate of the underlying autoregressive process by minimizing the prediction error vector $\mathbf{e} = \mathbf{x} - \mathbf{X}\mathbf{a}$ (commonly referred to as the residual):

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma\|\mathbf{a}\|_1, \tag{1}$$

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 $n < 1$

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.

and $n > N$ [8]. The more tractable 1-norm $\|\cdot\|_1$ is used here as a linear programming relaxation of the sparsity measure, often represented as the cardinality of a vector, i.e. 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 [9]. The choice of the regularization term γ is based on a trade-off between the sparsity of the residual and the sparsity of the predictor, found through by the L-curve [10]. The sparse structure of the predictor, allows a joint estimation of a short-term and a long-term predictors [7]:

$$A(z) \approx \hat{F}(z)\hat{P}(z) \tag{2}$$

where $\hat{F}(z)$ is the short-term predictor, commonly employed to remove short-term redundancies due to the formants, and $\hat{P}(z)$ is the long-term pitch predictor that removes the long-term redundancies. The two filters will then be quantized.

2.2. Step 2: Residual Estimation

In order to achieve frame independence, we rethink the analysis-by-synthesis (AbS) scheme used for the estimation of the approximated residual given $A(z)$, estimated in the previous step. In particular, the main equation of AbS coding is the following [3]:

$$\begin{aligned} \hat{\mathbf{r}} = \arg\min_{\mathbf{r}} &\|\mathbf{W}(\mathbf{x} - \hat{\mathbf{H}}\left[\hat{\mathbf{r}}_-^T,\, \mathbf{r}^T\right]^T)\|_2, \\ &\text{s.t.} \quad \textit{struct}(\mathbf{r}), \end{aligned} \tag{3}$$

where $\mathbf{x}$ is the $N \times 1$ frame of speech, $\mathbf{W}$ is the $N \times N$ perceptual weighting matrix, $\hat{\mathbf{H}}$ is the $N \times K + N$ synthesis matrix whose $i-$th row contains the elements with index $[0, K + i - 1]$ of the truncated impulse response $\hat{\mathbf{h}}$ of the combined quantized prediction filter $\hat{A}(z) = \hat{F}(z)\hat{P}(z)$:

$$\hat{\mathbf{H}} = \left[\begin{array}{ccc|ccccc} \hat{h}_K & \cdots & & \hat{h}_0 & 0 & 0 & \cdots & 0 \\ \hat{h}_{K+1} & \ddots & & \ddots & \ddots & 0 & 0 & 0 \\ \vdots & \ddots & & \ddots & \cdots & \hat{h}_0 & 0 & 0 \\ \hat{h}_{K+N-2} & \ddots & & \ddots & \cdots & \hat{h}_1 & \hat{h}_0 & 0 \\ \hat{h}_{K+N-1} & \hat{h}_{K+N-2} & & \cdots & \cdots & \hat{h}_2 & \hat{h}_1 & \hat{h}_0 \end{array}\right]. \tag{4}$$

The residual term $\left[\hat{\mathbf{r}}_-^T,\, \mathbf{r}^T\right]^T$ is composed of the K previous residual samples $\hat{\mathbf{r}}_-$ (the filter memory, already quantized) and the current $N \times 1$ residual vector $\mathbf{r}$ that has to be estimated. It is now clear that the dependence plays a central role in the estimation of the residual. The operator *struct*$(\cdot)$, that we will leave undefined at the moment, imposes the structure on the residual (e.g., MPE, RPE, CELP). Also, for the sake of simplicity, we will assume that no perceptual weighting is performed ($\mathbf{W} = \mathbf{I}$). The results can then be generalized for an arbitrary $\mathbf{W}$.

We now look for two estimates of the residual in (3), one where we take into consideration the previous residual $\hat{\mathbf{r}}_-$, one where we do not take it into consideration, therefore setting it to zero. The frame independent is then obtained considering only the $N \times N$ right side of the synthesis matrix in (4). The two residuals $\hat{\mathbf{r}}^{FI}$ and $\hat{\mathbf{r}}^{FD}$ will then be quantized.

2.3. Step 3: Re-estimation of the prediction coefficients

Once we have the two estimated residuals $\hat{\mathbf{r}}^{FI}$ and $\hat{\mathbf{r}}^{FD}$, we can calculate the truncated impulse response that generates them. In particular, we can rewrite the problem in (3) as:

$$\tilde{\mathbf{H}} = \arg\min_{\mathbf{H}} \|(\mathbf{x} - \mathbf{H}\hat{\mathbf{r}})\|_2 \rightarrow \tilde{\mathbf{h}} = \arg\min_{\mathbf{h}} \|(\mathbf{x} - \hat{\mathbf{R}}\mathbf{h})\|_2, \tag{5}$$

where

$$\hat{\mathbf{R}} = \left[\begin{array}{ccccccc} \hat{r}_0 & \cdots & \hat{r}_{-K} & 0 & 0 & \cdots & 0 \\ \hat{r}_1 & \ddots & \ddots & \ddots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \cdots & \ddots & 0 & 0 \\ \hat{r}_{N-1} & \ddots & \ddots & \cdots & \ddots & \hat{r}_{-K} & 0 \\ \hat{r}_N & \hat{r}_{N-1} & \cdots & \cdots & \cdots & \hat{r}_{-K+1} & \hat{r}_{-K} \end{array}\right], \tag{6}$$

is the $N \times N + K$ matrix constructed with the frame dependent residual vector $\left[\hat{\mathbf{r}}_-^T,\, \mathbf{r}^T\right]$. The problem (5) allows for a closed form solution when the 2-norm is employed in the minimization:

$$\tilde{\mathbf{h}} = \hat{\mathbf{R}}^T(\hat{\mathbf{R}}\hat{\mathbf{R}}^T)^{-1}\mathbf{x}, \tag{7}$$

with

$$\|\mathbf{x} - \hat{\mathbf{R}}\tilde{\mathbf{h}}\|_2 = 0. \tag{8}$$

We can now see that the optimal sparse linear predictor (frame dependent and frame independent) is the one that has $\tilde{\mathbf{h}}$ 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 as:

$$\hat{A}(z) = \hat{F}(z)\hat{P}(z) \rightarrow \hat{\mathbf{H}} = \hat{\mathbf{H}}_f\hat{\mathbf{H}}_p, \tag{9}$$

and re-estimate only the short-term impulse response, assuming that the long-term impulse response will not vary significantly, we can rewrite (5) using (9):

$$\tilde{\mathbf{h}}_f = \arg\min_{\mathbf{h}_f} \|(\mathbf{x} - \hat{\mathbf{H}}_p\hat{\mathbf{R}}\mathbf{h}_f)\|_2. \tag{10}$$

We can then obtain two estimates of the impulse responses, a frame dependent one $\tilde{\mathbf{h}}_f^{FD}$ and a frame independent one $\tilde{\mathbf{h}}_f^{FI}$. In the frame independent case, the matrix $\hat{\mathbf{R}}$ in (6) will be $N \times N$ and it will be constructed using only $\hat{\mathbf{r}}^{FI}$.

Using an autoregressive modeling of both $\tilde{\mathbf{h}}^{FD}$ and $\tilde{\mathbf{h}}^{FI}$, we obtain two new short-term predictive filters $\tilde{F}^{FI}(z)$ and $\tilde{F}^{FD}(z)$, that not only generate a better approximate of the impulse response but are also stable [4]. We will then quantize them.

2.4. Definition of an Enhancement Layer

For a given frame of speech we have calculated two residuals ($\hat{\mathbf{r}}^{FI}$ and $\hat{\mathbf{r}}^{FD}$) and two predictors ($\tilde{A}^{FI}(z) = \hat{P}(z)\tilde{F}^{FI}(z)$ and $\tilde{A}^{FD}(z) = \hat{P}(z)\tilde{F}^{FD}(z)$). The reconstructed speech frames are, for the frame independent case:

$$\hat{\mathbf{x}}^{FI} = \hat{\mathbf{H}}_p\tilde{\mathbf{H}}_f^{FI}\hat{\mathbf{r}}^{FI}, \tag{11}$$

and, for the frame dependent case:

$$\hat{\mathbf{x}}^{FD} = \hat{\mathbf{H}}_p\tilde{\mathbf{H}}_f^{FD}\left[(\hat{\mathbf{r}}_-^{FD})^T,\, (\hat{\mathbf{r}}^{FD})^T\right]^T. \tag{12}$$

It should be noted that $\hat{\mathbf{H}}_p$ is constructed from the truncated impulse response of $\hat{P}(z)$, that is equal for both cases, but in the frame independent case $\hat{\mathbf{H}}_p$ is $N \times N$ while in the frame dependent case $\hat{\mathbf{H}}_p$ is $N \times N + K$.

What we will do is transmit the frame independent parameters ($\hat{\mathbf{r}}^{FI}$, $\tilde{A}^{FI}(z) = \hat{P}(z)\tilde{F}^{FI}(z)$) to robustly construct a frame independent coder then define an enhancement layer based on the frame dependent parameters. To do so, we transmit the differences between

the two short-term predictors $\tilde{F}^{\Delta}(z)$ and the differences between the two residuals $\hat{\mathbf{r}}^{\Delta}(z)$. We will specify in the next section how to code the differences and in which domain.

If there is no loss of speech packets, it is clear that the decoder will work in "full" mode, using the frame independent informations together with the enhancement layer, (12) would then become:

$$\hat{\mathbf{x}} = \hat{\mathbf{H}}_p(\tilde{\mathbf{H}}_f^{FI} + \tilde{\mathbf{H}}_f^{EN}) \left[(\hat{\mathbf{r}}_-^{FI} + \hat{\mathbf{r}}_-^{EN})^T, (\hat{\mathbf{r}}^{FI} + \hat{\mathbf{r}}^{EN})^T \right]^T, \quad (13)$$

where $\tilde{\mathbf{H}}^{EN}$, $\hat{\mathbf{r}}_-^{EN}$ and $\hat{\mathbf{r}}^{EN}$ are functions of the parameters used to define the enhancement layer $\tilde{F}^{\Delta}(z)$ and $\hat{\mathbf{r}}^{\Delta}(z)$.

The interesting case is when a $k-$th frame is missing. In this case, the $k+1-$th frame is self-constructed only from the frame independent parameters, using (11). The $k+2-$th frame will then be reconstructed using the frame dependent information but first it is necessary to convert the part of the residual of the $k+1-$th frame $\hat{\mathbf{r}}_-^{FI}$, that will appear in the reconstruction equation (13), into the frame dependent one $(\hat{\mathbf{r}}_-^{FI} + \hat{\mathbf{r}}_-^{FE})$.

3. EXPERIMENTAL ANALYSIS

3.1. Setup

Linear predictive analysis

The length of the analyzed speech frames in our scheme is $N = 160$ (20 ms). The order of the optimization problem in (1) is $K = 110$, meaning that we can cover accurately pitch delays in the interval $[N_{stp}+1, K-N_{stp}-1]$, including the usual range for the pitch frequency $[70\text{Hz}, 500\text{Hz}]$. This also means that the dependency from the previous frame is $K = 110$ residual samples. The linear prediction filters $F(z)$ and $P(z)$ are chosen as respectively of order $N_f = 12$ and $N_p = 1$. $F(z)$ is coded initially as an LSF vector with 26 bits (providing transparent coding) using the procedure in [11]. The pitch period is coded with 7 bits and the gain with 6 bits.

Coding of the residual

The residual coding of both $\hat{\mathbf{r}}^{FI}$ and $\hat{\mathbf{r}}^{FD}$ is implemented using an RPE procedure [12] with fixed shift equal to zero and a sample spacing $Q = 8$. The RPE procedure is slightly modified to have the first 8 pulses as nonzero (27 nonzero pulses in total). This guarantees, other than a full row rank of $\hat{\mathbf{R}}$, also a well conditioned problem in (10) in both the frame dependent, where $\hat{\mathbf{R}}$ is $N \times N + K$ and frame independent case, where $\hat{\mathbf{R}}$ is $N \times N$. $\hat{\mathbf{r}}^{FI}$ is calculated first, then we impose the same sign structure when calculating $\hat{\mathbf{r}}^{FD}$. The residuals are also quantized simultaneously with a 8-level uniform quantizer, the peak magnitude is encoded with 6 bits per frame and 1 bit per pulse is used to code the sign.

Re-estimation procedure

In the re-estimation procedure (10), we impose the constraint of having $h_f(0) = 1$, this is done to simplify the IIR modeling of $\mathbf{h}_f$, so that the filter has a unit numerator. The new short-term predictive filters are also coded as an LSF vector with 26 bits (providing transparent coding in both cases).

Coding of the Enahncement Layer

The difference vector $\tilde{F}^{\Delta}(z)$ is calculated between $\tilde{F}^{FD}(z)$ and $\tilde{F}^{FI}(z)$ in the quantized LSF domain. A 11 bits vector quantizer has proved to be sufficient to describe the difference between the two polynomial. In particular, the reconstructed polynomial (sum of $\tilde{F}^{FI}(z)$ and $\tilde{F}^{\Delta}(z)$ in the LSF domain) is going to fulfill the spectral transparency performances as $\tilde{F}^{FD}(z)$ does. As for the difference between the two residuals $\hat{\mathbf{r}}^{\Delta}(z)$, we will use 2 bits per pulse, sufficient to code the difference almost without distortion in the quantized domain. Each frame will then be coded with a total of 218 bits, 153 belonging to the frame independent part and 65 belonging to the frame dependent enhancement layer, generating a total bit rate of 10.9 kbps (7.65 kbps for the frame independent information and 3.25 kbps for the enhancement layer).

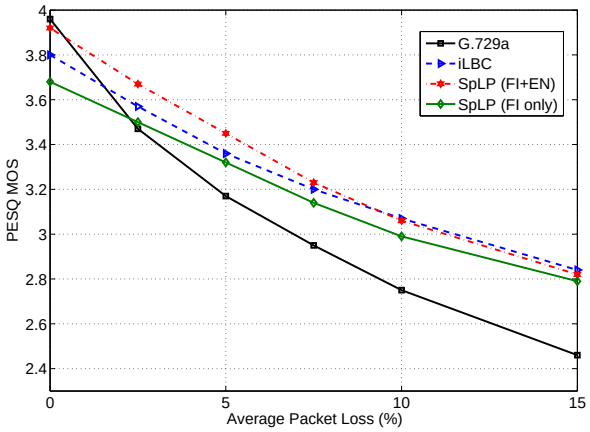

Fig. 1. Performances of the compared methods: G.729a (8 kbps), iLBC (13.33 kbps), and our introduced method based on sparse linear prediction (SpLP) with (FI+EN) and without (FI) the frame dependent enhancement layer (respectively 10.9 and 7.65 kbps).

3.2. Results

In this subsection we present the numerical results of our method compared, in terms of PESQ-MOS [13], to the iLBC in [1] and the G.729a [5], working respectively at 13.33 kbps and 8 kbps.

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 [14], re-sampled at 8 kHz. In our simulations, we used the Gilbert model for packet loss with parameters $q = P(\text{loss}|\text{loss}) = 0.7$ and $p = P(\text{loss}|\text{noloss})$ varied in order to have an average loss rate of $p/(p+q)$. The analyzed loss rates are 0%, 2.5%, 5%, 7.5%, 10%, and 15%. In our implementation, a simple packet loss concealment (PLC) based on repeating the previously received frames is implemented for our method and also for the G.729a.

As the results suggest in Figure 1, the coder works well with performances similar to the G.729a codec at 0% packet losses, where the iLBC fails to do so. The frame dependent layer seems to work well at low packet loss rates and loses its enhancement properties when the loss rate increases, as we may have expected. It should be noted, that our scheme, when only the frame independent part is employed, performs only slightly worse than iLBC with a net decrease in rate and a very simple PLC scheme. This can be explained by the novelty we have introuduced in the re-estimation of the "frame independent linear predictors" and by the compact and robust modeling advantages offered by sparse linear prediction [2]. Our coder performs worse than iLBC for loss percentage higher than 7.5%, mostly due to the more advanced PLC implemented on iLBC. A final comment is that the structured sparsity of the residual can allow guidance in order to generate an excitation sequence when packet loss occurs, for example when the other parameters are estimated in a Hidden Markov Model based PLC [15].

4. DISCUSSION

The coding algorithm we have presented is representative of a more general problem, where we minimize the expected distortion between the analyzed speech and its coded approximation, subject to a rate constraint:

$$\begin{array}{ll} \text{minimize} & D(\mathbf{x}, \hat{\mathbf{x}}), \\ \text{subject to:} & R(\hat{\mathbf{x}}) \leq R^*; \end{array} \tag{14}$$

where $D(\mathbf{x}, \hat{\mathbf{x}})$ represent the expected distortion by representing $\mathbf{x}$ with $\hat{\mathbf{x}}$, $R(\hat{\mathbf{x}})$ is the rate (or, equivalently, the bit allocation) to transmit $\hat{\mathbf{x}}$ and R^* is the maximum possible rate (the constraint). In our case, the distortion will be dependent on how the representation of $\hat{\mathbf{x}}$ divided between a frame independent core $\hat{\mathbf{x}}^{FI}$ and a frame dependent enhancement layer $\hat{\mathbf{x}}^{EN}$. In particular, the distortion term can be made dependent on the loss rate and therefore adjusting the bit allocation on the frame dependent and frame independent parts. We see for example from Figure 1 how the increase in performance given by the enhancement layer tend to reduce itself with the increase of the loss rate, in particular with a 15% of lost packets, there is almost no difference, although there is a 3.25 kbps difference in rate. In this case, what we would then like to do is to reallocate the bits used to define the enhancement layer, to improve the performances of the frame independent coder, the problem in (14) can then be rewritten as:

$$\begin{array}{ll} \text{min.} & w_{p_L} D(\mathbf{x}, \hat{\mathbf{x}}^{FI}) + (1 - w_{p_L}) D(\mathbf{x}, \hat{\mathbf{x}}^{FI} + \hat{\mathbf{x}}^{EN}), \\ \text{s.t.:} & R(\hat{\mathbf{x}}^{FI}) + R(\hat{\mathbf{x}}^{EN}) \leq R^*. \end{array} \tag{15}$$

where the allocation of the rate is now split between the frame independent part and the enhancement layer that exploits frame dependence. Also the expected distortion will be proportional to the different bit allocation. In (15), w_{p_L} is a weight that will be somehow proportional to the packet loss probability p_L ($0 \leq w_{p_L} < 1$), and, on a higher order analysis, it will also depend on other loss statistics such as the burst length. An interesting case, it is also to use the bit allocated for the enhancement layer to bring information for the packet loss concealment on how to reconstruct the missing frames when the loss rate is high. How to implement the problem in (15) will be subject of our future work.

5. CONCLUSION

In this paper, we have introduced a novel formulation for speech coding in packet networks. In particular, we have defined an algorithm that generates parameters that independently decode a speech segment at 7.65 kbps. A 3.25 kbps frame dependent enhancement layer is added to exploit inter-frame dependencies. This allows to reach performances similar to the G.729a coder for 0% packet loss probability while behaving similarly to the iLBC coder for higher packet loss probabilities. Sparse linear prediction has been used to robustly analyze a speech segment, providing a joint estimation of long-term and short-term predictors and a sparse residual. Also, a new formulation of the Analysis-by-Synthesis scheme has been defined by re-estimating a more appropriate synthesis matrix. A definition of the future work on the how to optimally construct a frame dependent/independent coder has also been given.

6. REFERENCES

[1] S. V. Andersen, et al., "iLBC - A linear predictive coder with robustness to packet loss", in *Proc. IEEE Workshop on Speech Coding*, pp. 23–25, 2002.

[2] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen "Speech coding based on sparse linear prediction", in *Proc. European Signal Processing Conference*, 2009.

[3] 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.

[4] D. Giacobello, M. N. Murthi, M. G. Christensen, S. H. Jensen, and M. Moonen, "Re-estimation of Linear Predictive Parameters in Sparse Linear Prediction", to appear in *Rec. 43rd Asilomar Conf. on Signals, Systems, and Computers*, 2009. Available online at `http://kom.aau.dk/~dg/publications.html`.

[5] ITU-T G.729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)", 2009.

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

[7] 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.

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

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

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

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

[12] 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.

[13] 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", 2001.

[14] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "DARPA-TIMIT acoustic-phonetic continuous speech corpus", *Technical Report NISTIR*, no. 4930, 1993.

[15] C. A. Rodbro, M. N. Murthi, S. V. Andersen, S. H. Jensen, "Hidden Markov Model-Based Packet Loss Concealment for Voice Over IP", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1609–1623, 2006.