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]**Aalborg Universitet, Aalborg, Denmark**
[2]**University of Miami, Florida, USA**
[3]**Katholieke Universiteit Leuven, Leuven, Belgium**

July, 2010

Motivation

- In VoIP systems, the most used approach is to create speech coders that are totally frame independent.
- In the case of telephony with dedicated circuits, high quality is achieved by the exploitation of inter-frame dependencies.
- Overcoming this mismatch by splitting 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.

Outline

1 Introduction

2 System Architecture

3 Validation

4 Conclusion

Prediction parameters estimation

- A sparse linear predictive framework is employed to achieve a more compact description of all the features extracted from a speech frame:

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

- The sparse structure of the high order predictor allows a joint estimation of a short-term and a long-term predictors $A(z) \approx \hat{F}(z)\hat{P}(z)$, the sparse residual allows for sparse multipulse encoding.

Residual Estimation

- We rethink the analysis-by-synthesis (AbS):

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

  where $\hat{\mathbf{H}}$ is the synthesis matrix obtained from the quantized prediction filter $\hat{A}(z) = \hat{F}(z)\hat{P}(z)$.
- 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.
- Find two residual estimates $\hat{\mathbf{r}}^{FD}$ (using $\hat{\mathbf{r}}_{-}$) and $\hat{\mathbf{r}}^{FI}$ (not using $\hat{\mathbf{r}}_{-}$).

Re-estimation of LP coefficients

- With $\hat{\mathbf{r}}^{FI}$ and $\hat{\mathbf{r}}^{FD}$, we calculate the truncated impulse response that generates them. In particular, we can rewrite the AbS equation as:

$$\tilde{\mathbf{h}} = \arg\min_{\mathbf{h}} \|(\mathbf{x} - \hat{\mathbf{R}}\mathbf{h})\|_2.$$

- We can split the two contribution as:

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

  and re-estimate only the short-term impulse response, assuming that the long-term impulse response will not vary significantly.
- We can then obtain two estimates of the impulse responses, a frame dependent one $\tilde{\mathbf{h}}_f^{FD}$ and a frame independent one $\tilde{\mathbf{h}}_f^{FI}$.
- 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.

Enhancement Layer

- The reconstructed speech frames are, for the frame independent case:

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

- and, for the frame dependent case:

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

- We transmit the frame independent parameters ($\hat{\mathbf{r}}^{FI}$, $\tilde{A}^{FI}(z) = \hat{P}(z)\tilde{F}^{FI}(z)$) and a side stream with the differences between the two short-term predictors $\tilde{F}^{\Delta}(z)$ and the differences between the two residuals $\hat{\mathbf{r}}^{\Delta}(z)$.

Working mode

- 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, would then become:

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

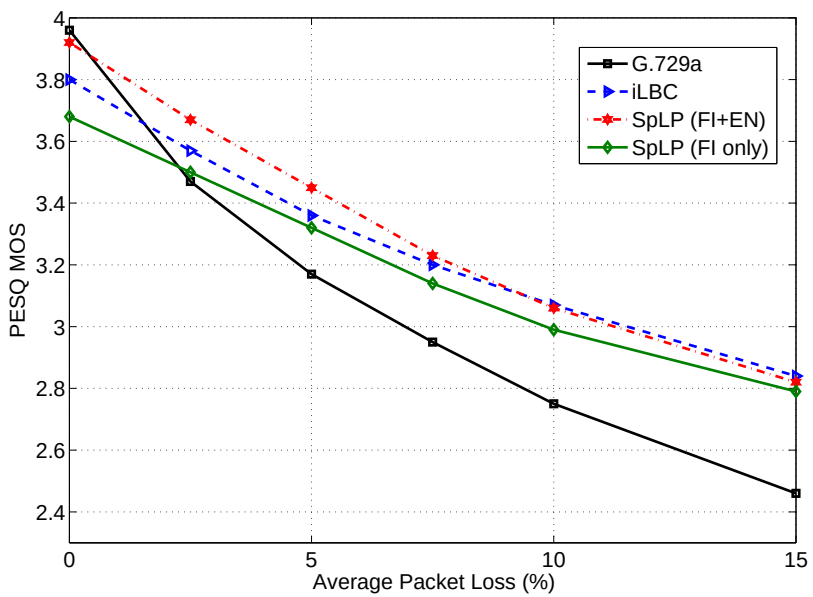
  where $\tilde{\mathbf{H}}^{EN}$, $\hat{\mathbf{r}}_-^{EN}$ and $\hat{\mathbf{r}}^{EN}$ are functions of the parameters used to define the enhancement layer $\tilde{F}^{\Delta}(z)$ and $\hat{\mathbf{r}}^{\Delta}(z)$.
- When a $k-$th frame is missing, the $k+1-$th frame is self-constructed only from the frame independent parameters. The $k+2-$th frame will 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, into the frame dependent one $(\hat{\mathbf{r}}_-^{FI} + \hat{\mathbf{r}}_-^{EN})$.

Experimental Setting

- The length of the analyzed speech frames in our scheme is $N = 160$ (20 ms). The order of the predictor $A(z)$ is $K = 110$. The linear prediction filters $F(z)$ and $P(z)$ are chosen as respectively of order $N_f = 12$ and $N_p = 1$.
- The residual coding of both $\hat{\mathbf{r}}^{FI}$ and $\hat{\mathbf{r}}^{FD}$ is implemented using an RPE procedure with fixed shift equal to zero and a sample spacing $Q = 8$.
- The difference vector $\tilde{F}^{\Delta}(z)$ is calculated between $\tilde{F}^{FD}(z)$ and $\tilde{F}^{FI}(z)$ in the quantized LSF domain.
- The difference between the two residuals $\hat{\mathbf{r}}^{\Delta}(z)$ will be coded with 2 bits per pulse, sufficient to code the difference almost without distortion in the quantized domain.

Results

- 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.



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).

Conclusions

- 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:

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

  where the allocation of the rate is now split between the frame independent part and the enhancement layer that exploits frame dependence.
- The expected distortion will be proportional to the different bit allocation (w_{p_L} proportional to packet loss percentage p_L ($0 \leq w_{p_L} < 1$) and burst length.
- The bit allocated for the enhancement layer can be also used to bring information for the packet loss concealment on how to reconstruct the missing frames when the loss rate is high.