

ENHANCEMENT OF CODED SPEECH BY CONSTRAINED OPTIMIZATION

W. Bastiaan Kleijn

Global IP Sound AB
Rosenlundsgatan 54
S-11863 Stockholm, Sweden

TMH Speech Signal Processing Group
KTH (Royal Institute of Technology)
S-10044 Stockholm, Sweden

ABSTRACT

A new method for the enhancement of speech signals contaminated by speech-correlated noise, such as that in the output of a speech coder, is presented. The method is based on constrained optimization of a criterion. We implement the method on a block-by-block basis and use two constraints. A first constraint ensures that the signal power is preserved. A modification constraint ensures that the power of the difference of the enhanced and unenhanced signal is less than a fraction of the power of the unenhanced signal. We apply the method to increase the periodicity of the speech signal. Noisy-sounding coded voiced speech generally has a high SNR and can be enhanced while satisfying a strict modification constraint. Sounds that are not nearly periodic are perceptually unaffected by the optimization because of the modification constraint. Practical results show that the method forms a powerful alternative to existing post-filter procedures.

1. INTRODUCTION

It is common practice to reduce audible speech-dependent (often called speech-correlated) noise in the output of speech-coding algorithms [1]. Such enhancement systems can be motivated using high-rate source-coding theory for stationary Gaussian processes with a mean-squared-error distortion criterion [2]. The power spectrum of the ideal reconstructed signal equals the power spectrum of the original signal minus the mean squared quantization error. This means that the decrease in the signal power spectrum is proportionally strongest in regions of low energy. In other words, the energy of the spectral valleys decreases proportionally more than that of spectral peaks, thus emphasizing the spectral shape. In speech-coding algorithms, the analysis and synthesis models are usually identical and quantization errors often lead to a deemphasis of the spectral shape. Thus, the results of source coding theory for Gaussian signals suggest an emphasis of the spectrum of the reconstructed signal by means of an adaptive post-filter. This qualitative statement remains correct if masking is considered.

For good performance, the coder and the adaptive post-filter are generally optimized as a complete system. Post-filtering, which was originally heuristically motivated, then leads to a coding structure that resembles the optimal coding structure for Gaussian signals. Post-filters for speech coders can be traced back to the work of Ramamoorthy and Jayant [3], who introduced an adaptive post-filter structure for the enhancement of coded speech. Chen and Gersho [4, 1] introduced the now ubiquitous adaptive pole-zero post-filter structure.

In general, the required emphasis of the spectrum can be performed separately for the spectral fine structure and for the spectral envelope. The spectral fine structure offers particular large poten-

tial for enhancement because of the large dynamic range of the harmonic structure of voiced speech. However, this potential for enhancement of the fine-structure is difficult to achieve because of implementation problems. Conventional adaptive post-filters are based on the coding parameters, and contain no feedback of the properties of the enhanced signal other than the signal power. This generally leads to a spectral emphasis that is too strong or too weak within different segments of a signal. Furthermore, the time synchronization between the spectral envelope and the spectral fine structure is generally incorrect in current fine-structure post-filters [5] because the inherent delay of the post-filter is ignored.

In the present paper, we propose a robust speech-enhancement procedure to reduce speech-dependent noise based on constrained optimization. The new technique avoids the problems of current post-filters in the enhancement of spectral fine-structure of speech.

2. CONSTRAINED ENHANCEMENT

Emphasis of the coded speech spectrum should lead to signal enhancement for almost all currently-used coding structures since they employ identical analysis and synthesis models and suffer from spectral deemphasis resulting from quantization. Emphasis of the spectrum can be achieved by constrained optimization of suitable measures. In particular, the decreased periodicity observed in coded speech can be mitigated by constrained optimization of a periodicity measure.

2.1. The Constraints

Let \mathbf{x}_j be a discrete speech segment of dimension K , with time label j . That is, \mathbf{x}_j is a sample sequence consisting of K subsequent speech samples. We will assume that K is sufficiently large that averaging over K is meaningful. Furthermore, let $\tilde{\mathbf{x}}_j$ be the sample sequence that replaces \mathbf{x}_j upon enhancement. It is reasonable and customary to make sure that the Euclidean norm of \mathbf{x}_j is preserved in the enhancement procedure:

$$\|\tilde{\mathbf{x}}_j\| = \|\mathbf{x}_j\|, \quad (1)$$

where $\|\cdot\|$ denotes the Euclidean norm. In our optimization-based enhancement procedure, the signal-norm preservation condition (1) becomes a first constraint.

The signal-norm constraint corresponds to the energy correction made in existing post-filtering procedures. Our optimization-based enhancement procedure makes the introduction of additional constraints natural. In particular, we can introduce a second constraint that the difference between \mathbf{x}_j and $\tilde{\mathbf{x}}_j$ is relatively small:

$$\|\mathbf{x}_j - \tilde{\mathbf{x}}_j\|^2 \leq \beta \|\mathbf{x}_j\|^2, \quad (2)$$

where $\beta \in [0, 1]$. We refer to this inequality constraint as the modification constraint. The modification constraint prevents that the enhancement procedure modifies the signal more than desirable.

In the present paper, we enhance the fine-structure of the speech signal. That is, we perform a constrained maximization of a periodicity measure. It is interesting to consider the effect of the constraints qualitatively. For voiced speech, the signal-norm constraint leads to a simultaneous reduction of energy in the spectral valleys between the speech harmonics and increase in energy of the spectral peaks (harmonics). Since the signal in the valleys has low energy per definition, the modification constraint either is not active or it affects performance of the enhancement procedure little. Thus, the audible enhancement of periodicity is strong. For signal segments that are not nearly periodic, the modification constraint prevents a change in the perceived quality of the signal.

2.2. Periodicity Maximization

Let $\mathbf{x}_{j,m}$ denote a sample sequence that contains K samples that are each m pitch cycles removed from the corresponding sample of the sequence $\mathbf{x}_j = \mathbf{x}_{j,0}$. (Note that $\mathbf{x}_{j,m}$ and $\mathbf{x}_{j,m+1}$ can overlap.) We can define as a measure of periodicity of the enhanced signal

$$\eta_{\mathcal{J}} = \sum_{j \in \mathcal{J}} \sum_{m \in \mathcal{I} - \{0\}} \alpha_m \langle \tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_{j,m} \rangle, \quad (3)$$

where α_m describes a discrete window function, \mathcal{I} is a set of integers that describes the support of this window (e.g., $\mathcal{I} = \{-3, -2, \dots, 3\}$), and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product ($\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|^2$) and \mathcal{J} is a set of consecutive-block indices. The window $\{\alpha_m\}_{m \in \mathcal{I}}$ should be defined based on perception. We can maximize criterion (3) by iteratively maximizing the criteria

$$\eta_j = \sum_{m \in \mathcal{I} - \{0\}} \alpha_m \langle \tilde{\mathbf{x}}_j, \mathbf{x}_{j,m} \rangle. \quad (4)$$

Good results can be obtained with a simplified procedure. We use one iteration for each η_j based on the original $\mathbf{x}_{j,m}$ and apply the constraints to the individual optimizations.

We first consider the case where the modification constraint is an equality constraint. Based on equation (4) and constraints (1) and (2), we define the extended criterion

$$\eta'_j = \sum_{m \in \mathcal{I} - \{0\}} \alpha_m \langle \tilde{\mathbf{x}}_j, \mathbf{x}_{j,m} \rangle + \lambda_1 \|\tilde{\mathbf{x}}_j\|^2 + \lambda_2 \|\mathbf{x}_j - \tilde{\mathbf{x}}_j\|^2, \quad (5)$$

where λ_1 and λ_2 are Lagrange multipliers.

It is convenient to define the sample sequence \mathbf{y}_j :

$$\mathbf{y}_j = \sum_{m \in \mathcal{I} - \{0\}} \alpha_m \mathbf{x}_{j,m}. \quad (6)$$

The extremum for $\tilde{\mathbf{x}}_j$ of the extended criterion then satisfies

$$0 = \frac{\delta \eta'_j}{\delta \tilde{\mathbf{x}}_j} = \mathbf{y}_j + 2\lambda_1 \tilde{\mathbf{x}}_j + 2\lambda_2 (\mathbf{x}_j - \tilde{\mathbf{x}}_j). \quad (7)$$

If we write

$$\tilde{\mathbf{x}}_j = A\mathbf{y}_j + (B+1)\mathbf{x}_j \quad (8)$$

we find, after some algebra

$$A = \left(\frac{(\beta - \frac{\beta^2}{4}) \|\mathbf{x}_j\|^2}{\|\mathbf{y}_j\|^2 - \frac{\langle \mathbf{y}_j, \mathbf{x}_j \rangle}{\|\mathbf{x}_j\|^2}} \right)^{\frac{1}{2}} \quad (9)$$

and

$$B = -\frac{\beta}{2} - A \frac{\langle \mathbf{y}_j, \mathbf{x}_j \rangle}{\|\mathbf{x}_j\|^2}. \quad (10)$$

Next, we consider the case where the modification constraint (2) is a true inequality. That means that the modification constraint is not activated, and we can set $\lambda_2 = 0$ in equation (5):

$$\eta'_j = \sum_{m \in \mathcal{I} - \{0\}} \alpha_m \langle \tilde{\mathbf{x}}_j, \mathbf{x}_{j,m} \rangle + \lambda_1 \|\tilde{\mathbf{x}}_j\|^2. \quad (11)$$

In this case, the optimization results in

$$\tilde{\mathbf{x}}_j = C\mathbf{y}_j, \quad (12)$$

with

$$C = \sqrt{\frac{\|\mathbf{x}_j\|^2}{\|\mathbf{y}_j\|^2}}, \quad (13)$$

that is, $\tilde{\mathbf{x}}_j$ equals \mathbf{y}_j , scaled to preserve the signal norm.

It is now straightforward to define a general algorithm for obtaining the enhanced sample sequence $\tilde{\mathbf{x}}_j$ by maximizing the periodicity subject to both the signal-norm constraint and the modification constraint. The algorithm is applied to subsequent sample sequences. We first compute the optimal solution for $\tilde{\mathbf{x}}_j$ ignoring the modification constraint and check whether this solution satisfies the modification constraint. If it does not satisfy this constraint, we compute and use the solution for the case where the modification constraint is an equality constraint. The algorithm can be summarized as

1. compute \mathbf{y}_j , A , B , and C ,
2. if $\|\mathbf{x}_j - C\mathbf{y}_j\|^2 \leq \beta \|\mathbf{x}_j\|^2$
then $\tilde{\mathbf{x}}_j = C\mathbf{y}_j$
else $\tilde{\mathbf{x}}_j = A\mathbf{y}_j + (B+1)\mathbf{x}_j$.

2.3. Practical Aspects: Determination of $\{\mathbf{x}_{j,m}\}_{m \in \mathcal{I}}$

Thus-far, we have tacitly assumed that the pitch-synchronous set of sample sequences $\{\mathbf{x}_{j,m}\}_{m \in \mathcal{I}}$ is available for a given sample sequence \mathbf{x}_j . (We use sample sequences of dimension $K = 40$ to $K = 80$. Within this range, the performance of the enhancement procedure does not depend strongly on the value of K .)

In the implementation, we determine the sequences $\mathbf{x}_{j,m}$ sequentially, starting from \mathbf{x}_j . In the backward direction we first determine $\mathbf{x}_{j,-1}$, then $\mathbf{x}_{j,-2}$ based on $\mathbf{x}_{j,-1}$, and so on. In the forward direction we first determine $\mathbf{x}_{j,1}$, then $\mathbf{x}_{j,2}$ based on $\mathbf{x}_{j,1}$, etc. We constrain the delay between all corresponding samples of sequences $\mathbf{x}_{j,m}$ and $\mathbf{x}_{j,m+1}$ to be the same.

Let $\mathbf{w}_{j,m+1}(t)$ denote a candidate sequence for $\mathbf{x}_{j,m+1}$ located at time separation t from $\mathbf{x}_{j,m}$. (Note that $\mathbf{w}_{j,m+1}(t)$ is a K -dimensional vector, not a sample value, and that t can be noninteger.) We then determine the correlation sequence $\langle \mathbf{x}_{j,m}, \mathbf{w}_{j,m+1}(t) \rangle$ ranging over a set of t values. We select t such that

$$\langle \mathbf{x}_{j,m}, \mathbf{w}_{j,m+1}(t) \rangle \geq \langle \mathbf{x}_{j,m}, \mathbf{w}_{j,m+1}(t') \rangle, \quad \forall t' \in \mathcal{A}, \quad (14)$$

where \mathcal{A} defines the allowed set of time locations of $\mathbf{x}_{j,m+1}$ and $t \in \mathcal{A}$. In our implementation, we select the set \mathcal{A} as a small set of values based on the output of a pitch estimator or based on an adaptive-codebook delay. The method is robust against errors in the selection of the set \mathcal{A} , because of the modification constraint.

3. PERFORMANCE AND DISCUSSION

Figure 1 illustrates the operation of the constrained periodicity-enhancement procedure. For this example, we set $\beta = 0.05$, which is a typical value for noisy-sounding coded speech. This value of β corresponds to a signal to modification power ratio of about 13 dB. The enhancement procedure is operating at all time and does not have any information about whether the signal is voiced or unvoiced. The figure shows that, for voiced speech, the audible noise present in the valleys between the signal harmonics is reduced by the enhancement procedure. This is possible because the signal-to-noise ratio is more than 13 dB. On the other hand, the method does not change the reconstructed-signal quality for unvoiced speech because the modification is no more than 13 dB. The continuously changing pitch track also implies that the periodicity is not enhanced during signal regions that do not contain nearly periodic signals.

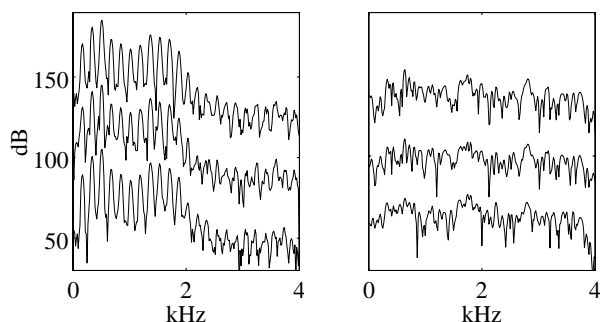


Fig. 1. Illustration of the operation of the enhancer with a maximum modification ratio of 13 dB. On the left power spectra of voiced speech and on the right power spectra of unvoiced speech. Lower spectra are original signals, middle spectra are coded signals, and top spectra are enhanced coded signals.

The perceptual quality resulting from the constrained periodicity enhancement procedure was evaluated with formal and informal testing procedures. In general, for signals with a noisy or rough character, the enhancement obtained from the procedure is immediately clear even from casual listening. The method is also capable of enhancing signals with very high signal-to-noise ratios. Informal tests show that the low-level audible noise that can be heard in a signal encoded with the ITU G.711 standard can be removed with the proposed constrained enhancement procedure.

The constrained periodicity enhancement procedure was formally tested on a real-time waveform coder operating on 8 kHz sampled speech. The sequence dimension was $K = 80$ and initial estimates for the delay set \mathcal{A} were based on delays obtained by the adaptive-codebook procedure [6] used in the waveform coder. For $\{a_m\}_{m \in \mathcal{I}}$ a Hanning window with seven-sample support was used. The forward part of this window was cut where necessary to maintain a delay of less than 10 ms. This delay is comparable with the implicit time-varying delay of conventional pitch post-filters (this delay is generally ignored). Thus, for short pitch periods, seven pitch cycles are considered simultaneously and this number is reduced for long pitch periods. The resulting improvement in the formal MOS testing is shown in table 1. The performance of the enhancement operation did not display a significant dependency on gender or speaker identity.

Naturally, if the modification constraint is relatively loose (that is, β is relatively large), the periodicity becomes too strong

Table 1. MOS for various conditions. The standard error of the reported MOS values is about 0.075.

condition	MOS
MNRU 24 dB	3.32
MNRU 30 dB	4.06
coder	3.52
coder with enhancer, $\beta = 0.05$	3.86

in the high frequency bands. This effect can be mitigated by decomposing each K -dimensional sequence $\mathbf{x}_{j,m}$ by means of a filterbank. There is no advantage to approximating ideal filters but it is advantageous to use perfect-reconstruction filterbanks. We found that a simple three-times over-sampled three-band perfect-reconstruction filterbank corresponding to a uniform filterbank frame [7] based on the orthonormal vectors $\frac{1}{2}[1, \sqrt{2}, 1]$, $\frac{1}{2}[1, -\sqrt{2}, 1]$, and $\frac{1}{\sqrt{2}}[1, 0, -1]$ facilitates the effective removal of visibly over-emphasized periodicity in the high-frequency band of the power spectrum. However, informal tests showed no audible advantage for the decomposition in the normal operating range of the enhancement procedure and the decomposition was not used to obtain the MOS test results.

4. CONCLUSIONS

We introduced a method for enhancing coded speech based on constrained optimization. The method is very effective at both low and high signal-to-noise ratios and forms an alternative to conventional post-filtering. The method is inherently robust because it cannot introduce large changes to the signal. The described implementation of the method is sufficiently powerful that post-filtering of the spectral envelope can be avoided. This means that the fidelity of the spectral envelope is maintained, thus facilitating tandeming. An audio demonstration of the method can be found at <http://www.speech.kth.se/~bastiaan/enhancer.html>.

5. REFERENCES

- [1] J. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 59–71, 1995.
- [2] S. V. Andersen and W. B. Kleijn, "Reverse water-filling in predictive encoding of speech," in *IEEE Speech Coding Workshop*, (Porvoo), pp. 105–107, 1999.
- [3] V. Ramamoorthy and N. S. Jayant, "Enhancement of ADPCM speech by adaptive postfiltering," *AT&T Bell Labs. Tech. J.*, pp. 1465–1475, 1984.
- [4] J. Chen and A. Gersho, "Real-time vector APC speech coding at 4800 bps with adaptive postfiltering," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, (Dallas), pp. 2185–2188, 1987.
- [5] W. B. Kleijn, "Improved pitch prediction," in *Proc. IEEE Workshop on Speech Coding for Telecomm.*, (Sainte-Adele, Quebec), pp. 19–20, 1993.
- [6] W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, "Analysis and improvement of the vector quantization in SELP," in *Proc. Mobile Satellite Conference*, (Pasadena), pp. 527–532, 1988.
- [7] H. Bölcskei, F. Hlawatsch, and H. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Proc.*, vol. 46, no. 12, pp. 3256–3268, 1998.