PRE-PROCESSING OF NOISY SPEECH FOR VOICE CODERS

Taran Agarwal and Peter Kabal

Department of Electrical & Computer Engineering, McGill University
3480 University Street, Montreal, Quebec, Canada, H3A 2A7

ABSTRACT

Accurate Linear Prediction Coefficient (LPC) estimation is one of the key requirements for low bit-rate voice coding. Under harsh acoustic conditions, LPC estimation can become unreliable. This results in poor quality of encoded speech and introduces annoying artifacts.

This paper presents a two-branch speech enhancement preprocessing scheme for low bit-rate voice coders. This scheme consists of two parallel denoising blocks. One block will enhance the degraded speech for LPC estimation. Another block will increase the perceptual quality of the speech to be coded. The goal of this paper is to design the two-branch scheme. Test results show that the two-branch scheme can provide better perceptual quality compared to conventional one-branch speech enhancement techniques in noisy environments.

1. INTRODUCTION

In recent years, considerable progress has been achieved in reducing the bit-rate while maintaining a high level of speech quality. Although vocoders, such as ITU G.729 and Mixed Excited Linear Prediction (MELP), give high quality for clean speech, it is significantly worse for coded noisy speech. One solution to circumvent this issue is to add a speech enhancement pre-processor that attenuates noise in the corrupted speech prior to encoding. Although several denoising algorithms exist, see [1], and may be used as front-end processors, there is a need for application-specific speech enhancement.

A typical vocoder relies heavily on accurate LPC estimation [2]. Under noisy conditions, the LPC estimation is disturbed. In 1999, Martin et al., derived an algorithm that was optimized for LPC estimation [3]. In this paper their algorithm will be referred to as MMSE Adaptive Limiting Scheme for Estimation (MMSE-ALSE) estimator.

In the same year, Accardi et al., proposed the use of two parallel denoising algorithms as a pre-processing stage (Fig. 1) prior to low bit-rate coding [4]. The goal of such a modular pre-processing approach is to have one denoising block (referred to as ‘Type L’ in Fig. 1) targeted at processing speech for improved LPC estimation, while another block for computation of the residual signal (referred to as ‘Type R’ in Fig. 1).

Since MMSE-ALSE is already “optimized” for LPC estimation, it is of interest in this work to define another denoising algorithm aimed at improving the perception of reconstructed speech. The derived denoising algorithm will be used for ‘Type R’, while MMSE-ALSE for ‘Type L’ enhancement [5]. The derived speech enhancement is built on the existing MMSE-ALSE estimator described in [6, 7].

This paper is organized as follows: Section 2 summarizes the parameters used by the MELP speech coder; Section 3 takes a cursory look on the MMSE-LSA algorithm and its importance in denoising; Section 4 introduces the Perceptual Evaluation of Speech Quality (PESQ); Section 5 explains the procedure adopted to derive the proposed denoising algorithm; Section 6 presents the results of listening tests and objective measures as suggested in [8].

2. PARAMETERS OF THE MELP SPEECH CODER

Traditional vocoders use either periodic pulses or white noise as the excitation for a synthesis filter. Most of these vocoders produce intelligible speech at very low bit-rates, but they often sound synthetic and are prone to occasional annoying tonal thumps and buzzing. Since these problems stem from the inability of the periodic pulses to mimic all kinds of voiced speech, MELP uses both, a mixture of pulse and noise excitation. The model for MELP uses a mixture of lowpass filtered pulse train and highpass filtered noise, with the mixture strength controlled by an analysis of the bandpass voicing strengths [2]. In 1996, the US DoD selected MELP as a new federal standard. It is used as a testbed in our two-branch pre-processor. The 2400 bps MELP coder extracts 1 pitch value, 5 bandpass voicing strength values, 1 aperiodic/periodic flag, 2 gain factors, 10 Fourier coefficients and 10 LP coefficients from an input speech frame of 180 samples.

3. MMSE-LSA ESTIMATOR

The MMSE-LSA speech enhancement algorithm consists of three stages: spectral analysis/synthesis (through windowed FFT/IFFT and over-lap add), noise Power Spectral Density estimation (periodogram or exponential averaging over silence), and a spectral gain computation [5]. The close relation of the MMSE-LSA estimator to the Itakura-Saito measure, its ability to reduce the annoying effects of musical noise (see [9]), and its straightforward parameterization on the a priori and a posteriori SNR [5], make it a suitable algorithm to build on.

This work was supported by an NSERC / Nortel Networks Industrial Research Chair.

Fig. 1. Two-branch pre-processor scheme and a basic parametric voice coder.

---

0-7803-7549-1/02/$17.00 ©2002 IEEE.
The MMSE-LSA estimator minimizes $E\{ (\log \hat{A}_k - \log A_k)^2 \}$ where $A_k = |S_k|$ is the spectral speech amplitude of the $k$th spectral bin and $\hat{A}_k$ is the best estimate of speech corrupted with noise: $Y_k = S_k + N_k$, where $N_k$ is additive noise. $A_k$ is obtained by multiplying $Y_k$ with $G_{\text{LSA}}(\xi_k, \gamma_k)$:

$$G_{\text{LSA}}(\xi_k, \gamma_k) = \frac{\xi_k}{1 + \xi_k} \exp \left\{ \frac{1}{2} \int_{v_k}^{\infty} e^{\frac{-t}{\gamma_k}} dt \right\} ,$$ (1)

where $\xi_k$ and $\gamma_k$ are interpreted as the conditional a priori SNR and a posteriori SNR respectively and $v_k = \xi_k^{-\gamma_k} / (1 + \xi_k)$ [5]. $\xi_k$ conditioned on the presence of speech: $\xi_k = \eta_k / (1 - q_k)$, where $q_k$ is the probability of speech absence and $\eta_k$ is the unconditional a priori SNR obtained using the ‘decision-directed’ approach [10]. Accardi et al., showed that further improvements in the estimator can be obtained by incorporating a multiplicative modifier [7]:

$$G_{\text{MM}}(\xi_k, \gamma_k, q_k) = \frac{\mu_k}{\mu_k + (1 + \xi_k) \exp(-v_k)} ,$$ (2)

where $\mu_k = (1 - q_k) / q_k$. The total gain:

$$G_{\text{TOT}}(\xi_k, \gamma_k) = G_{\text{LSA}} \cdot G_{\text{MM}} ,$$ (3)

is multiplied with $Y_k$ to obtain the estimate $\hat{A}_k$.

In 1999, Cox et al., proposed using Eq. (4) as an adaptive lower limit to improve LPC estimation [3]:

$$\eta_{\text{min}} = \sqrt{\text{SNR} - 16.5} ,$$ (4)

where SNR is the input speech SNR (in dB) and $\eta_{\text{min}}$ (in dB) is the adaptive lower limit on $\eta_k$. This adaptive limit is only applied to signal frames, while a constant $\eta_{\text{min}} = 0.12$ is applied to noise only frames. The resulting time-varying $\eta_{\text{min}}$ is recursively smoothed with smoothing parameter of $\alpha = 0.8$.

4. EVALUATION OF SPEECH QUALITY

The cost-function used in the derivation of the new adaptive lower limit on $\eta_k$ is the output obtained from the PESQ algorithm—released by ITU-T as P.862. The PESQ algorithm takes two inputs: uncoded and coded speech, and gives the Mean Opinion Score (MOS) for the coded speech. The output has shown to have a correlation coefficient of 0.935 with 22 known ITU benchmark puts: uncoded and coded speech, and gives the Mean Opinion limit on $\eta$. The cost-function used in the derivation of the new adaptive lower limit on $\eta_k$ is the spectral speech amplitude of the $k$th spectral bin and $G_{\text{TOT}}(\xi_k, \gamma_k)$ for 12 fixed arbitrary limits on $\eta_k$ (referred to as $\eta_{\text{min}}(R)$), ranging from $-45$ to $-3$ dB. As an example, consider a female speech at $0$ dB. This file was enhanced 12 times. Each file was MELP coded and then processed with the PESQ algorithm to obtain 12 MOS ratings. These MOS ratings were plotted against their corresponding $\eta_{\text{min}}(R)$ and interpolated using a cubic spline. The $\eta_{\text{min}}(R)$ that corresponds to the maximum MOS rating was recorded and plotted against 0 dB (see Fig. 2). Similarly other speech files were processed to obtain the result seen in Fig. 2. The cubic interpolation line of best fit is:

$$\eta_{\text{min}}(R) = 0.0013(\text{SNR})^3 - 0.1(\text{SNR})^2 + 2.5(\text{SNR}) - 38 ,$$ (5)

where $\eta_{\text{min}}(R)$ (in dB) is the newly derived adaptive lower limit on $\eta_k$ to be used in Eq. (3) and SNR is the input speech SNR in dB. Unlike Eq. (4), $\eta_{\text{min}}(R)$ is applied to all frames irrespective of it being speech or silence. The new noise suppression algorithm will be referred to as MMSE-Adaptive Limiting Scheme for Perception (MMSE-ALSP).

6. RESULTS

Using some of the objective measures suggested in [8], it was noticed that MMSE-ALSP gave the best results for LPC estimation, while MMSE-ALSP gave the highest percentage of correct pitch prediction even under harsh acoustic conditions ($\approx 4\%$ at an SNR of 0 dB), see Fig. 3 and Fig. 4. These results were obtained by corrupting 12 s of male speech with synthetic white noise and using MMSE-ALSP, MMSE-ALSE, the Enhanced Variable Rate Coder noise suppression (EVRCS) and MMSE-LSA algorithms.

Following A–B subjective comparison tests for several combinations of Type L and Type R enhancement algorithms, two schemes emerged as the most preferred pre-processing schemes for Fig. 1 and are listed in Table 1.

For the selected schemes in Table 1 more results were generated with: babble, music, Hoth and car noise under various acous-
Table 1. Selected pre-processing schemes.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Enhancement Algorithm</th>
<th>Type L</th>
<th>Type R</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>MMSE-ALSE</td>
<td>MMSE-ALSP</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>MMSE-ALSE</td>
<td>MMSE-LSA</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Minimum Euclidean distance between LSF parameters.

Fig. 4. Percentage correct pitch prediction.

tic environments\(^2\) (e.g., 5, 10, 20 dB), see Fig. 5. From Fig. 5 it is evident that Scheme I can be used as a pre-processor for low-bit rate coders even under harsh acoustic conditions for several noisy environments (≈ 0.1 MOS improvement in babble environment at an SNR of 5 dB).

7. CONCLUSION

In this paper the problem of degraded speech quality of vocoders in the presence of background noise was addressed. An algorithm that was built on the existing MMSE-LSA estimator was introduced that aims at improving the perceptual quality of encoded speech. It is shown that a two-branch pre-processor scheme can give better auditory impression of speech coded at very low-bit rates.

8. REFERENCES


\(^2\)Test data was prepared according to Supplement 23 to ITU-T P-series Recommendations [13].

Fig. 5. MOS under various acoustic conditions for several noise types, as obtained by the PESQ algorithm.