Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Signal Processing 86 (2006) 1089–1093

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Perceptual weighting filter for robust speech modification

Joon-Hyuk Chang

Department of Electronic Engineering, Inha University, Incheon, 402-751, Korea

Received 18 February 2005; received in revised form 12 May 2005; accepted 28 July 2005
Available online 30 August 2005

Abstract

In this paper, an improved preprocessor for low-bit-rate speech coding employing the perceptual weighting filter is proposed. Speech modification in the proposed approach is performed according to a criterion which makes a compromise between the modification and perceptual weighted quantization errors. For this, the perceptual weighting filter is expressed in terms of a transform domain matrix. The proposed approach is effective in enhancing the speech signal at coder-decoder (CODEC) output through a number of listening tests.

Keywords: Speech modification; Perceptual weighting filter; Quantization error

1. Introduction

In general, the performance of a low-bit-rate speech coder degrades seriously under the presence of various interfering signals such as background noise, acoustic echo, music sounds or interfering speaker's speech. This phenomenon is mainly due to the deviation from the assumed speech production model which is used in the codebook training since a number of codebooks used in the coder are trained based on a large amount of speech data and the ranges for parameter search are specified to fit the pure speech signals. One of the successful applications of the unwanted distortion reduction technique to low-bit-rate coding is the speech enhancement technique [1–3,8]. Even though aforementioned enhancement techniques have been found effective in the presence of a stationary background noise, they are not capable of handling such interfering signals as the acoustic echoes, music sounds or co-talkers' speech. This is mainly due to the fact that the conventional approaches adopt the open loop analysis which cannot take advantage of speech coder characteristics. An alternative method is the generalized *analysis-by-synthesis* (AbS) technique where the original input speech signal is modified such that it can be coded more efficiently [5]. Recently, a preprocessor is developed to modify the signal applied to a certain speech coder based on a system identification problem [4]. A distinguished feature of this algorithm is that a criterion which compromises between the modification error and the quantization

E-mail address: changjh@hi.snu.ac.kr.

0165-1684/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2005.07.025

error has been employed successfully in the objective function for the quality improvement.

In this paper, we propose a perceptually improved preprocessor which incorporates the perceptual weighting filter for the less audible quantization error. The perceptual weighting filter is expressed as the transform domain matrix and is also being combined with the constant modification factor which differs from our previous work [4] in that a separate modification factor is assigned to each frequency bin. The principal advantage of this method is the fact that the perceptual weighting filter brings us strong modification on formant valleys in which audible quantization noise mainly exists. From a number of experiments, the presented improved preprocessor has been found to improve the perceived speech quality compared with the original preprocessor when the background interfering signal is added.

2. System identification for system matrix

We first briefly review the basic theory for system matrix estimation [4]. Using the generalized AbS paradigm, the input speech signal is modified before being fed to the coder so that it can be reconstructed in the receiver side with minimal distortion. In the presented approach, prior to applying to the encoder, we modify $\mathbf{x}$ such that the modified vector can better fit to the speech coder. Let $\mathbf{y} = [y(0), y(1), \ldots, y(M-1)]^{\mathrm{T}}$ be the signal samples obtained by modifying $\mathbf{x}$, and $\mathbf{z} = [z(0), z(1), \ldots, z(M-1)]^{\mathrm{T}}$ be the output vector which is produced when $\mathbf{y}$ is applied to the coder and then re-synthesized in the decoder. Also let $\mathbf{Y} = [Y(0), Y(1), \ldots, Y(N-1)]^{\mathrm{T}}$ and $\mathbf{Z} = [Z(0), Z(1), \ldots, Z(N-1)]^{\mathrm{T}}$ be the transform domain representation of $\mathbf{y}$ and $\mathbf{z}$, respectively. Without loss of generality, we assume that $\mathbf{Z} = \mathbf{Q}(\mathbf{Y}_a)$ where $\mathbf{Y}_a^{\mathrm{T}} = [\mathbf{Y}_p^{\mathrm{T}} \,|\, \mathbf{Y}^{\mathrm{T}} \,|\, \mathbf{Y}_f^{\mathrm{T}}]$ is an augmented input vector, and $\mathbf{Q}(\cdot)$ represents the transfer function that models the input–output characteristic of the CODEC. Also, $\mathbf{Y}$ represents the input data on the current frame. On the other hand, $\mathbf{Y}_p$ stands for the previous data and $\mathbf{Y}_f$ consists of the future input samples which are usually referred to as the *look-ahead* data. On the other hand, since the CODEC output $\mathbf{Z}$ is mostly affected by the current input $\mathbf{Y}$, the effects of the previous and look-ahead date are assumed to be ignored without a significant modeling error. For that reason, it can be assumed that $\mathbf{Z} = \mathbf{Q}(\mathbf{Y})$.

Estimation of the transfer matrix $\mathbf{Q}$ is obtained from choosing a system identification technique. To give a derivation for estimation of the transfer matrix $\mathbf{Q}$, we follow the recursive least square (RLS) estimation procedure proposed in [4] which gives a more detailed descriptions for the estimation of the transfer matrix $\mathbf{Q}$.

3. Perceptual weighting filter for speech modification

Usually, the perceptual weighting procedure often results in improvement in the speech coder performance. A commonly used weighting filter is based on the linear prediction (LP) coefficients that represent the short-term correlation in the speech signal [6]. A representative perceptual weighting filter $W(z)$ is given by

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1 - \sum_{i=1}^{p} a_i z^{-i}}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}}, \tag{1}$$

where $A(z)$ represents the pth-order LP analysis filter and a_i is the LP coefficient. To compute the filter coefficients for this filter, linear predictive analysis is used [6]. Also, γ is a perceptually weighting factor which does not alter the center formant frequency, but just broadens the bandwidth of the formants. Specifically, frequency broadening δf given by $\delta f = (f_s/\pi) \ln \gamma\,(\mathrm{Hz})$, where f_s is the sampling frequency in hertz. For that reason, the weighting filter deemphasizes the formant structure while emphasizing the formant valleys of the speech signal. This results in a larger matching error in the region of the formants, where spectral masking makes the auditory systems less sensitive to quantization error. The most suitable value of γ is selected subjectively by listening tests, and for 8 kHz sampling, γ is adopted as 0.9 here.

If we assume $W(z)\ (= 1 - \sum_{k=1}^{p} w_k z^{-k})$ in z domain has a time domain response $f(n)$ which is an finite impulse response (FIR) sequence of the form, $f(n)$ is given by

$$f(n) = \begin{cases} 1, & n = 0, \\ -w(n), & 1 \leqslant n \leqslant p, \\ 0 & \text{otherwise}, \end{cases}$$

where we chose $p = 12$ which is the experimentally selected value. In other words, we truncate $f(n)$ only to the first 12 samples although the impulse response lasts longer. As a consequence, we can obtain comfortable fit to a vowel segment and the smoothed spectrum which is considered to be suitable in terms of the subjective speech quality.

Hence we can calculate $W(e^{j\omega})$, using the discrete Fourier transform (DFT), by supplementing $f(n)$ with sufficient zero-valued samples to form N-point sequence. We take the DFT of the zero-padded $f(n)$ sequence giving $W(e^{j(2\pi/N)k})$, $0 \leqslant k \leqslant N-1$. Given $\mathbf{Q}$, modification of the input vector $\mathbf{X}$ is achieved according to the following criterion: $\hat{\mathbf{Y}} = \arg\min_{\mathbf{Y}} J(\mathbf{Y})$. Although the identity matrix is given for the perceptual weighting filter in [4], we briefly review the criterion for speech modification such that

$$J(\mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|_{\mathbf{W}}^2 + K\|\mathbf{Y} - \mathbf{QY}\|_{\mathbf{W}}^2, \tag{2}$$

where K is called the (positive constant) modification factor for speech modification and $\mathbf{W}$ is the perceptual weighting filter.[1] Here, we present the new objective function incorporating the perceptual weighting filter for the quantization error such that

$$J(\mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|^2 + K\|\mathbf{Y} - \mathbf{QY}\|_{\mathbf{W}_Q}^2, \tag{3}$$

where $\mathbf{W}_Q$ denotes the diagonal matrix of the perceptual weighting filter for the quantization errors, respectively. Letting us differentiate $J(\mathbf{y})$ with respect to $\mathbf{y}$ then

$$\frac{\partial J(\mathbf{Y})}{\partial \mathbf{Y}} = -2(\mathbf{X} - \mathbf{Y}) + 2K\tilde{\mathbf{Q}}^{\#}\mathbf{W}_Q\tilde{\mathbf{Q}}\mathbf{Y}, \tag{4}$$

where $\tilde{\mathbf{Q}} = \mathbf{I} - \mathbf{Q}$ with $\mathbf{I}$ being the $N \times N$ identity matrix. Also, # means the Hermitian operations.

Equating it to zero, (4) can be written as

$$\hat{\mathbf{Y}} = [I + K\tilde{\mathbf{Q}}^{\#}\mathbf{W}_Q\tilde{\mathbf{Q}}]^{-1}\mathbf{X} \tag{5}$$

$$= [\mathbf{I} + \mathbf{G}_w\tilde{\mathbf{Q}}^{\#}\tilde{\mathbf{Q}}]^{-1}\mathbf{X}. \tag{6}$$

From (6), it is observed that $\mathbf{G}_w$ replaces $K\mathbf{W}_Q$ in the proposed approach. Summarizing this derivation, $\mathbf{G}_w$ becomes the perceptual weighted modification factor which means that a separate speech modification factor is assigned to each frequency bin, which is considered more robust and realistic.

From (2), it is apparent that the amount of modification and quantization errors are controlled by the positive constant modification factor K. If K is large, more emphasis is placed on the quantization error and a larger modification of the input speech is allowed. However, in [4], a fixed K may be equally applied to all frequency components which cannot take full advantage of the perceptual weighting principle. In formant valleys of the given speech signal, the disturbing quantization noise can be even more disturbing to a human listener. This requires a necessity for a strong modification on the formant valleys. As we mentioned, the formant valleys are emphasized while the formant parts are deemphasized by the perceptual weighting filter. On the other hand, since $\mathbf{G}_w$ is a value which the positive constant K is incorporated with the perceptual weighting filter, the higher modification can be applied to the formant valleys while the lower modification may be applied to the formant regions. This choice has advantages for strong modification over a formant null where the audible quantization noise is mainly located. Through a number of listening tests, it has been found that the perceptual weighting filter renders the minimally audible quantization noise at the expense of an increase in complexity due to the LP analysis.

Fig. 1 shows a typical example of the magnitude spectra of the clean, noisy and modified input speech when we focus on the voiced sounds part. We also depict the corresponding modification factor which is a higher value at the formant valleys in Fig. 1. According to the figure, we can see that the proposed approach strongly modifies the formant valleys compared with the original method.

[1] $\|\mathbf{a}\|_W^2 = \mathbf{a}^{\#}W\mathbf{a}$.

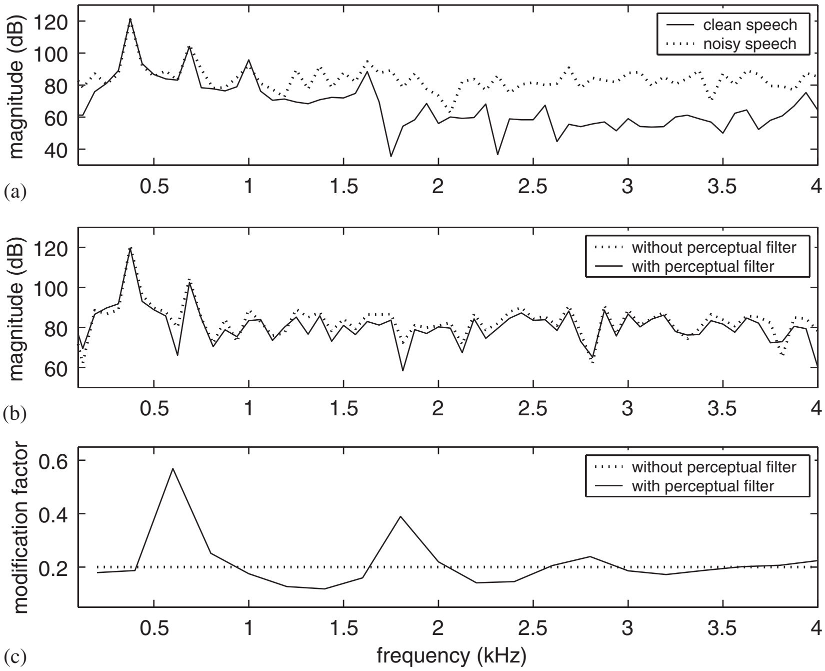

Fig. 1. Magnitude spectra and perceptual weighted modification factor for a voiced sounds: (a) clean and noisy speech; (b) modified speech and (c) modification factor.

4. Experimental results

The subjective quality of the proposed approach was evaluated using the mean opinion score (MOS) for a wide variety of 10 listeners. Twenty test sentences, in which ten were generated by a male speaker and the others by a female speaker, were used for quality measurement. Each sentence was sampled at 8 kHz, and the frame size was 10 ms. As a target speech coder, we employ the ITU-T 8 kb/s speech coder G.729A [8]. For input speech modification, each frame of data was transformed into a vector consisting of the corresponding DFT coefficients, and the modification was done in the 80 point DFT domain.

To simulate the noisy environments, we added the white and babble noises from the NOISEX-92 database by varying signal-to-noise ratio (SNR). Moreover, the background music signal and co-talker's speech were also used to degrade the input speech quality. The MOS results are shown in Table 1 where all the scores were obtained with $K = 0.2$. From the results, we can see that the proposed approach gives us improved results compared with the original method in most of the tested conditions. Performance improvement was found greater for the white and music environments compared to other cases. In the case of the interfering speech, the proposed method yielded nearly same performance in most of the tested SNR conditions.

Table 1
MOS results with 95% confidence intervals

Condition (SNR)	G.729A	Original [4]	Proposed
Clean speech	4.12 ± 0.02	4.14 ± 0.02	4.14 ± 0.02
White noise (5 dB)	2.00 ± 0.07	2.18 ± 0.07	2.30 ± 0.07
White noise (10 dB)	2.75 ± 0.08	2.90 ± 0.08	3.11 ± 0.09
Babble noise (5 dB)	2.70 ± 0.09	2.90 ± 0.10	2.95 ± 0.10
Babble noise (10 dB)	3.07 ± 0.10	3.17 ± 0.10	3.27 ± 0.10
Music (5 dB)	2.80 ± 0.11	3.12 ± 0.11	3.25 ± 0.11
Music (10 dB)	3.20 ± 0.12	3.35 ± 0.12	3.50 ± 0.11
Interfering speech (5 dB)	2.52 ± 0.10	2.80 ± 0.10	2.85 ± 0.10
Interfering speech (10 dB)	3.02 ± 0.10	3.15 ± 0.11	3.17 ± 0.11

5. Conclusions

We have proposed an approach to input speech modification to be used by designing a preprocessor based on the perceptual weighting filter. Based on the simplified system modeling of the given codec transfer function, the objective function of the optimization problem is described as a compromise between the modification and perceptually weighted quantization error. From a number of experiments, the proposed perceptually improved speech modification has been found beneficial in reducing the quantization error and superior to the original modification technique.

Acknowledgements

The author would like to thank four anonymous reviewers for invaluable suggestions.

References

[1] A.J. Accardi, R.V. Cox, A modular approach to speech enhancement with an application to speech coding, in: Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing, vol. 1, March 1999, pp. 201–204.

[2] J.-H. Chang, N.S. Kim, Speech enhancement: new approaches to soft decision, IEICE Transactions and Information and Systems, vol. 27(E84-D), September 2001, pp. 1231–1240.

[3] H. Da, Z. Cao, A joint speech coding-enhancement algorithm for MBE vocoder, in: Proceedings of International Conference on Communication Technology, vol. 2, October 1998, pp. 22–24.

[4] N.S. Kim, J.-H. Chang, A preprocessor for low-bit-rate speech coding, IEEE Signal Processing Letters, vol. 9(10), October 2002, pp. 318–321.

[5] W.B. Kleijn, R.P. Ramachandran, P. Kroon, Interpolation of the pitch-predictor parameters in analysis-by-synthesis coders, IEEE Transactions on Speech and Audio Processing, vol. 2(1), January 1994, pp. 42–54.

[6] W.B. Kleijn, K.K. Paliwal, Speech Coding and Synthesis, Elsevier Science B.V., Amsterdam, 1995.

[8] R. Salami, et al., Design and description of CS-ACELP: a toll quality 8 kb/s speech coder, IEEE Transactions on Speech and Audio Processing, vol. 6(2), March 1998, pp. 116–130.