Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (Orlando, FL), pp. I-537-I-540, May 2002

SIGNAL SUBSPACE SPEECH ENHANCEMENT WITH PERCEPTUAL POST-FILTERING

Mark Klein and Peter Kabal

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

ABSTRACT

A methodology for suppressing musical noise produced by signal subspace speech enhancement is presented. An auditory post-filter is placed at the output of the subspace filter to smooth the enhanced speech spectra. By utilizing a perceptual filter, averaging is performed in a manner similar to that of the human auditory system. As such, distortion to the underlying speech signal is reduced.

1. INTRODUCTION

In most speech enhancement systems, *musical noise* can be attributed to errors in measuring noise statistics. This auditory annoyance resembles a sum of sinusoids of changing frequencies, turning "off" and "on" over successive frames. Signal subspace techniques eliminate musical noise originating from fluctuating energy estimates by averaging over long windows. However, other artefact sources exist. These include rapid changes of model order and subspace swapping. The latter condition refers to noise basis vectors being incorrectly employed to describe the signal subspace.

This paper presents a methodology to quell artefacts produced by signal subspace techniques. A perceptual post-filter is placed at the output of the signal subspace filters to smooth the enhanced signal. It will be shown that psychoacoustic knowledge can attenuate imperfections with minimal distortion to the speech signal being recovered.

Perception has been employed to the speech enhancement problem on several occasions. In [1, 2, 3], it was shown that the utilization of properties of the human auditory system has the capability to attenuate noise without distortion. Jabloun showed in [4] that knowledge of the ear can improve parameter estimates for signal subspace techniques. In this work, filter coefficients are derived using eigenvalues which are calculated by projecting the excitation pattern of the noisy signal onto the squared magnitude of the individual eigenvectors.

It is the goal of the perceptual post-filter to remove all traces of musical noise. Its strengths are two-fold: (1) distortion is minimized by attenuating only what is audible, and (2) peaks within the noise residual are smoothed by spectral and temporal averaging. However, the underlying speech should not be affected.

Limiting the attenuation in an enhancement scheme can decrease distortion. In this application, the perceptual filter accomplishes this by attenuating artefacts until they lie close to the masking threshold. As such, some of the artefact which is imperceptible is retained. By attenuating less, it is expected that fewer disturbances will be produced.

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

Spectral averaging increases the width of tones within the noise residual according to the resolution of the ear. Temporal averaging, by limiting magnitude changes of the noise residual over several frames, effectively attenuates musical noise. Rapid frame-to-frame spectrum variations are with high probability, the product of noise. By considering human perception, artefacts can be smoothed without noticeably altering the underlying speech signal.

This paper will possess the following structure: Section 2 will introduce the concept of signal subspace speech enhancement techniques. The ideas of subspace decomposition and linear estimation will be discussed. Section 3 will describe the operation of the perceptual post-filter. Results and discussion will be presented in Section 4. Finally, Section 5 will provide conclusions.

2. SIGNAL SUBSPACE BASED SPEECH ENHANCEMENT

Signal subspace-based speech enhancement techniques decompose M-dimensional spaces into two subspaces: a signal subspace and a noise subspace. It is assumed that the speech signal can lie only within the signal subspace while the noise spans the entire space. Thus, only the components of the signal subspace are used to estimate the original speech signal. This paper will employ the Karhunen-Loève expansion method for signal subspace speech enhancement originally proposed by Ephraim and Van Trees in [5].

2.1. Problem Formulation

The speech enhancement problem shall be described as a speech signal $\boldsymbol{x}$ being transmitted through a distortionless channel that is corrupted by additive white noise $\boldsymbol{w}$ with variance σ_w^2. The resulting output noisy speech signal $\boldsymbol{y}$ can be expressed as

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w} \tag{1}$$

where $\boldsymbol{x} = [x_1, x_2, \ldots, x_M]^T$, $\boldsymbol{w} = [w_1, w_2, \ldots, w_M]^T$ and $\boldsymbol{y} = [y_1, y_2, \ldots, y_M]^T$. The observation period has been denoted as M.

The speech enhancement system will attempt to estimate the original signal using a single channel of received speech. If the noise signal is not white, a prewhitening filter, $\boldsymbol{R}_w^{-0.5}$, may be applied to $\boldsymbol{y}$. Henceforth, $\boldsymbol{R}_{(\cdot)}$ will denote the correlation matrix of a signal.

2.2. Karhunen-Loève Expansion Subspace Decomposition

Performing an eigendecomposition on the speech signal correlation matrix, the following expansion is obtained

$$\boldsymbol{R}_x = \boldsymbol{Q}\boldsymbol{\Lambda}_x\boldsymbol{Q}^H = \begin{bmatrix}\boldsymbol{Q}_1 & \boldsymbol{Q}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_{x_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{Q}_1^H \\ \boldsymbol{Q}_2^H\end{bmatrix} \tag{2}$$

where $\mathbf{\Lambda}_{x_1} = \mathrm{diag}(\lambda_{x_1}, \ldots, \lambda_{x_K})$ and K denotes the number of non-zero eigenvalues of $\boldsymbol{R}_x$.

The eigenvector matrix $\boldsymbol{Q}$ has been partitioned into two submatrices, $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$. The matrix $\boldsymbol{Q}_1$ contains eigenvectors with corresponding non-zero eigenvalues. These eigenvectors form a basis for the signal subspace. Meanwhile, $\boldsymbol{Q}_2$ contains the eigenvectors which span the noise subspace.

The speech correlation matrix and the dimension of $\boldsymbol{Q}_1$ must be estimated from noisy data. Intuitively, the correlation matrix of the original speech signal can be estimated from the noisy correlation matrix by

$$\boldsymbol{R}_x = \boldsymbol{R}_y - \sigma_w^2 \boldsymbol{I}. \tag{3}$$

The dimension of the signal subspace can be approximated from the number of strictly positive eigenvalues for $\boldsymbol{R}_x$. Therefore,

$$K^* = \#\{k \in \mathbb{Z}_+ : \lambda_{x_i} > 0\} \tag{4}$$

The projection of the noisy speech onto the signal subspace can be obtained by applying operator $\boldsymbol{Q}_1\boldsymbol{Q}_1^H$ to $\boldsymbol{y}$. However, noise will still exist within the signal subspace. Thus, additional steps will be taken to better approximate the original speech signal.

2.3. Spectral Domain Constraint Signal Estimators

Once the noise subspace has been removed, a linear estimator is applied. This attenuates the noise in the signal subspace producing a better approximation of the original signal $\boldsymbol{x}$.

The filter matrix for the linear estimator shall henceforth be denoted as $\boldsymbol{H}$. The speech estimate resulting from applying the estimator can be calculated as $\hat{\boldsymbol{x}} = \boldsymbol{H}\boldsymbol{y}$.

For the ensuing section, the residual of the clean signal, $\boldsymbol{e}$, can be represented as

$$\boldsymbol{e} = \hat{\boldsymbol{x}} - \boldsymbol{x} = (\boldsymbol{H} - \boldsymbol{I})\boldsymbol{x} + \boldsymbol{H}\boldsymbol{w} = \boldsymbol{e}_x + \boldsymbol{e}_w \tag{5}$$

where $\boldsymbol{e}_x$ will refer to the *signal distortion* and $\boldsymbol{e}_w$ will denote the *residual noise*.

The spectral domain constraint estimator attempts to minimize the distortion to the speech signal while constraining noise below a threshold. This optimization can be expressed in the following manner:

$$\begin{aligned} \boldsymbol{H}^* &= \arg\min_{\boldsymbol{H}} \mathrm{tr}\{E\{\boldsymbol{e}_x\boldsymbol{e}_x^H\}\} \\ \text{subject to: } & \begin{array}{ll} E\{|\boldsymbol{q}_i^H\boldsymbol{e}_w|^2\} \le \alpha_i\sigma_w^2 & i = 1, \ldots, K \\ E\{|\boldsymbol{q}_i^H\boldsymbol{e}_w|^2\} = 0 & i = K{+}1, \ldots, M \end{array} \end{aligned} \tag{6}$$

where $0 \le \alpha \le 1$. From Eq. (6), the optimal filter is found to be

$$\boldsymbol{H}^* = \boldsymbol{Q}_1(\Lambda_{x_1}(\Lambda_{x_1} + \sigma_w^2\boldsymbol{I})^{-1})\boldsymbol{Q}_1^H \tag{7}$$

The speech signal resulting from the estimator given in Eq. (7) will contain noticeable musical noise. Application of the perceptual post-filter will mitigate the problem without adding significant distortion.

3. OVERVIEW OF THE ENHANCED SIGNAL SUBSPACE METHOD

The signal subspace filter will be modified to suppress musical noise by appending a perceptual post-filter to the output of the signal subspace filter. It should be stressed that this filter does not significantly attenuate the noise. Rather, it smoothes its input in a manner that musical noise is diminished and speech is unaffected.

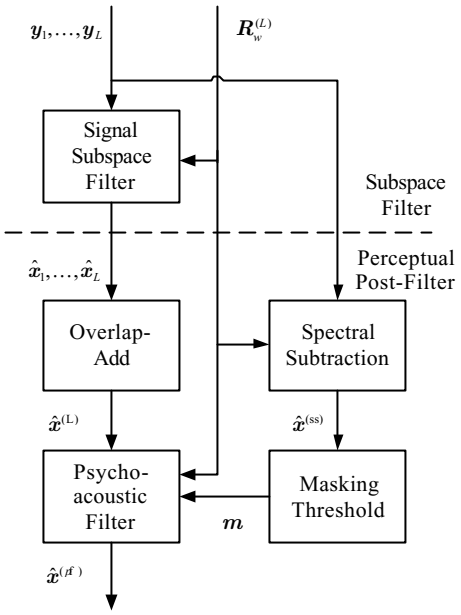


Fig. 1. Flow Chart of the Improved Signal Subspace Enhancement Scheme

A flow-chart describing the operation of the modified speech enhancement scheme can be found in Fig. 1.

The signal subspace filter operates most effectively when utilizing very short frames ($<$ 15 ms). Unfortunately, such frames do not provide sufficient frequency resolution for the calculation of a masking threshold. Thus, L input frames are sent to the signal subspace filter ($\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L$). The outputs, $\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_L$, are later merged, thereby increasing the frequency resolution of the speech estimate.

The psychoacoustic filter attempts to conceal the salient noise using the perceptual properties of the ear while minimizing the distortion to the underlying speech. This block is signal dependent, requiring an estimate of the noise correlation matrix, $\boldsymbol{R}_w^{(L)}$ and the masking threshold of the speech signal, $\boldsymbol{m}$, to calculate an appropriate gain.

The input to the psychoacoustic filter is $\hat{\boldsymbol{x}}^{(L)}$, the concatenation of L output frames from the signal subspace filter, $\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_L$. The frames are combined by the overlap-add block which utilizes appropriate windows and overlap length.

The masking threshold is calculated utilizing the model described in the Perceived Audio Quality ITU Recommendation (ITU-R BS.1387) [6]. For use in the perceptual post-filter, the masking model has been modified to operate with arbitrary sampling frequencies and time windows. As the clean speech signal is unavailable, it is necessary to estimate the masking threshold of the speech signal from noisy data. Thus, the spectra of the clean speech is estimated using the spectral subtraction technique.

3.1. Description of the Psychoacoustic Filter

The auditory filter minimizes signal distortion while constraining the spectrum of the noise residual to be beneath the masking threshold. The constrained optimization problem can be summa-

rized as

$$\boldsymbol{T}^* = \arg\min_{\boldsymbol{T}} \; E\{\|(\boldsymbol{T}\boldsymbol{F}^H - \boldsymbol{F}^H)\hat{\boldsymbol{x}}^{(L)}\|_2^2\}$$
$$\text{subject to: } E\{|\boldsymbol{f}_i^H \boldsymbol{T}\boldsymbol{w}^{(L)}|^2\} \leq m_i^2 \quad (8)$$

where $\boldsymbol{T}^* = \text{diag}\{t_1^*, \ldots, t_N^*\}$. N denotes the concatenated frame size and $\boldsymbol{F}$ is the Fourier transform matrix.

The optimal filter can be formulated as

$$t_i^* = \begin{cases} 1 & m_i \geq (s_{w_i}^{(L)})^{\frac{1}{2}} \\ \dfrac{m_i}{(s_{w_i}^{(L)})^{\frac{1}{2}}} & m_i < (s_{w_i}^{(L)})^{\frac{1}{2}} \end{cases}. \quad (9)$$

where $\boldsymbol{s}_w^{(L)}$ is the power spectral density of the noise signal.

Intuitively, it may seem more appropriate to attempt to place the noise residual of the signal subspace filter beneath the masking threshold. However, it was determined empirically that this criteria added musical noise due to the peakiness of the filter transfer function.

4. DISCUSSION AND RESULTS

The implementation issues of the enhanced signal subspace algorithm will first be discussed. Afterwards, the results of several experiments will explored. Finally, several design issues will be addressed.

4.1. Implementation

The signal subspace filter has been implemented with a frame size of 100 taps and 50% overlap. The psychoacoustic filter utilized a 300 sample frame length. A rectangular analysis window is applied to the data prior to signal subspace filtering. After application of the post-filter, a sine-squared synthesis window is utilized for reconstruction.

The 100 lags of the signal correlation matrices are calculated using 350 samples multiplied by a rectangular window. The noise covariance matrix is updated during speech pauses. As the noise signal is assumed to have slowly varying statistics, this method should be sufficient to obtain accurate estimates.

In all simulations, an 8 kHz sampling rate is used. SNR is calculated using the approach in ITU recommendation ITU-T P.830 [7]. The SNR is defined to be the ratio of the active speech level [8] to the RMS noise level.

4.2. Evaluation of the Efficacy of the Perceptual Post-filter

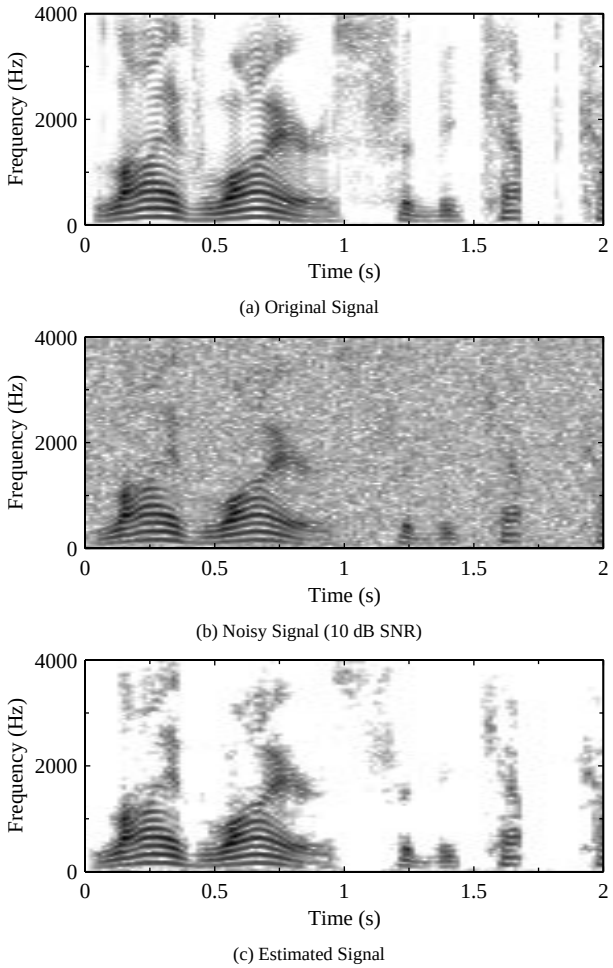
The efficacy of the perceptual post-filter was determined by comparing the outputs of the speech enhancement algorithm with the perceptual post-filter absent or present.

Informal listening tests determined that without the perceptual post-filter the speech contained noticeable musical noise. When the perceptual post-filter was applied, these artefacts are largely attenuated, although, slight distortion was noticeable in the enhanced speech. These observations were made with speech corrupted by white noise with 10 dB SNR.

A spectrogram of the sentence "Live wires should be kept covered" is depicted in Fig. 2. It shows a speech file before and after enhancement with the proposed algorithm, as well as clean state. The speech file has been perturbed by white noise and has an SNR of 10 dB.

It is evident from the denoised signal that the noise has been mostly removed. Unlike many enhancement algorithms which tend to muffle speech, this method retains the high frequency components. This ensured that the enhanced signal possessed naturalness.

It was also noted that voiced speech was better handled than unvoiced speech. This was attributable to the innate suitability of voiced speech for low-rank representations. In contrast, unvoiced signals require near-full-rank models to be modelled accurately. Accordingly, a subspace reduction will produce better results for voiced phonemes.



(a) Original Signal

(b) Noisy Signal (10 dB SNR)

(c) Estimated Signal

Fig. 2. Comparison of Speech Spectrograms

It should be emphasized that the psychoacoustic filter does not affect the underlying speech. Trials where high SNR input was passed through the perceptual post-filter produced perceptually unchanged outputs.

4.3. Performance in Varied Adverse Environments

The performance of the system was examined under several different operating conditions. The SNR was assigned values of 6 dB, 10 dB, and 15 dB with additive white noise. Coloured noise was then presented to the enhancement scheme to determine its flexibility with different noisy environments.

When the high SNR signals of 15 dB were processed, the speech signal was recovered without distortion or artefacts. For an SNR of 10 dB, it was observed that the only voiced speech was

recovered without error. For unvoiced speech, the weakest sounds suffered the most distortion. Overall, the fidelity of the recovered speech was still quite high. A loss of naturalness was detected with the 6 dB SNR sentences. Intelligibility, however, remained quite high for all test files.

The proposed algorithm was also tested with several environmental noise files, including car noise and pink noise. It was found that full speech recovery was possible with engine noise. In contrast, the enhanced signal arising from pink noise displayed significant musical noise. This was due to the fact that these disturbances were produced by poor noise estimation. When the true noise correlation matrices were utilized, the artefacts abated. In all cases, subjects preferred the enhanced speech to the noisy speech.

4.4. Comparison with Spectral Subtraction

The enhanced signal subspace method was compared with the spectral subtraction method. The evaluation was performed utilizing test data corrupted by additive white noise with 10 dB SNR. The spectral subtraction utilized an oversubtraction factor to improve performance. A window size of 256 samples with 50% overlap was employed. To ensure that the spectral subtraction algorithm did not suffer from poor noise estimation, the instantaneous spectra of the noise signal was made available.

Preliminary listening tests have shown that the enhanced signal subspace method is preferable to the spectral subtraction algorithm. A sizeable oversubtraction factor had to be used with the spectral subtraction method to give musical noise suppression comparable to the proposed algorithm. Accordingly, noticeable distortion was produced with the traditional method.

4.5. Examination of Computational Complexity

The signal subspace filter and perceptual post-filter had computational complexities of $O(n^3)$ and $O(n^2)$ respectively. However, the complexity of these blocks may be reduced by using simpler models. The computational requirements of the signal subspace filter could have been increased by utilizing non-optimal transforms. Furthermore, the post-filter could have been simplified by employing less sophisticated masking models than PEAQ. For this work, the best transform and masking model available were chosen to prove the efficacy of the proposed algorithm.

4.6. Additional Design Issues

The spectral domain constraint estimator and order estimators were modified to try to improve the fidelity of the enhanced speech. In the course of experimentation, an exponent was used in the gain of the Wiener filter as follows

$$\boldsymbol{H}^* = \boldsymbol{Q}_1(\Lambda_{x_1}(\Lambda_{x_1} + \sigma_w^2\boldsymbol{I})^{-1})^{\gamma}\boldsymbol{Q}_1^H. \qquad (10)$$

As the exponent, γ, is increased, the transition from the low SNR gain of 0 to the high SNR gain of 1 is sharpened. However, it was observed that modifying γ in the range of [0.5,1.5] did not produce audible benefits.

Several order estimators were also considered for measuring subspace dimensionality. The MDL order estimator proposed by Wax and Kailath [9] based on the work of Rissanen was evaluated. However, it was found that this formulation was not robust in the simpler case of determining the number of sinusoids in additive white noise. Thus, it was considered to be unsuitable for the more difficult task of speech order estimation in adverse conditions.

The order estimator put forth by Merhav [10] was also examined but eventually put aside. This algorithm was employed in the original paper by Ephraim and Van Trees [5]. It was determined that the minimal improvement obtained in enhanced speech quality did not justify the significant computational complexity.

The simple estimator described in Section 2.3 was found to be the best compromise between computational complexity and performance.

5. CONCLUSION

In this work, a frame-work to attenuate musical noise produced by signal subspace speech enhancement methods was presented. This speech restoration system incorporates the auditory concept of masking to smooth spectral parameters. Through informal listening tests, it has been shown that this algorithm is effective at attenuating musical noise while leaving speech relatively undistorted. It has been further ascertained that the speech enhancement algorithm is well suited for many adverse noise environments. Finally, it was determined that the proposed method outperformed the spectral subtraction algorithm.

6. REFERENCES

[1] G. A. Soulodre, *Camera Noise from Film Soundtracks*. Ph.D. thesis, McGill University, Department of Electrical Engineering, Nov. 1998.

[2] N. Virag, "Signal channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 126–137, Mar. 1999.

[3] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech and Audio Processing*, vol. 5, pp. 497–514, Nov. 1997.

[4] F. Jabloun and B. Champagne, "On the use of masking properties of the human ear in the signal subspace speech enhancement approach," in *Int. Workshop on Acoustic Echo and Noise Control*, (Darmstadt, Germany), Sept. 2001.

[5] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 251–266, July 1995.

[6] *Method for objective measurements of perceived audio quality*, Recommendation ITU-R BS.1387, International Telecommunication Union, July 1999.

[7] *Methods for Objective and Subjective Assessment of Quality*, Recommendation ITU-T P.830, International Telecommunication Union, Feb. 1996.

[8] *Objective Measurement of Active Speech Level*, Recommendation ITU-T P.56, International Telecommunication Union, Mar. 1993.

[9] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 33, pp. 387–392, Apr. 1985.

[10] N. Merhav, "The estimation of model order in exponential families," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1109–1114, Sept. 1989.