PERCEPTUAL GAIN FUNCTION FOR EIGENSPECTRAL DOMAIN SPEECH ENHANCEMENT

Vinesh Bhunjun, Mike Brookes

Imperial College London, Dept. of Electrical Engineering, London SW7 2AZ, UK

ABSTRACT

The goals of speech enhancement are to improve its perceptual aspects most commonly by applying a gain function to the noisy signal coefficients in a transform domain. This gain function is normally chosen to provide a good trade-off between suppressing the noise and avoiding speech distortion. In this paper, we identify some desired gain characteristics for better sounding enhancement and propose a method for choosing the gain transfer function based on perceptual criteria. We implement our approach in the eigenspectral domain and compare our results with those from selected eigenspectral-based transfer functions.

Index Terms— Eigenspectral domain speech enhancement, Coloured noise, Perceptual-based gain derivation, Residual noise energy, Distortion noise energy

1. INTRODUCTION

Speech recorded in a noisy environment undergoes degradation that affects quality, in the form of an increased noise level, and impairs intelligibility of the speech signal. The goals of speech enhancement are to improve its perceptual aspects most commonly by applying a gain function to the noisy signal coefficients in a transform domain. This gain function is normally chosen to provide a good trade-off between suppressing the noise and avoiding speech distortion [1].

Ephraim and Van Trees (EVT) [2] propose an eigenspectral domain speech enhancement algorithm for speech corrupted with white noise. They express the estimation error energy as a sum of a noise-related component, called residual noise energy, and a speech-related one, called speech distortion energy, which they relate to the concepts of quality and intelligibility respectively. They derive a gain function by minimizing the speech distortion energy while constraining the residual noise energy to be below a certain level, giving the Time-Domain Constrained (TDC) estimator. Using an alternative constraint in which the spectrum of the residual noise is shaped to match that of the speech, they also derive the Spectral-Domain Constrained (SDC) estimator.

EVT [2] suggest prewhitening the noisy signal if the noise is not white as assumed in the TDC/SDC derivation. Mittal and Phamdo [3] point out that, with prewhitening, the constraints for the minimization are applied to the transformed noisy signal and not the original one. Rezayee and Gazor [4] observe that the noisy signal eigenvectors nearly diagonalize the noise covariance matrix and so they use a diagonal matrix of noise energy values. Hu and Loizou [5] simplify the SDC minimization using a matrix that jointly diagonalizes the noise and signal covariance matrices, an approach extended further by Lev-Ari and Ephraim [6].

EVT [2] do not indicate how to specify the constraints on the residual noise spectrum. Hu and Loizou [7] adapt a technique used in low-rate speech coding for allowing a higher level of quantization noise near peaks in the power spectrum of the speech since the noise is masked by the speech. They define a perceptual filter that weighs the noise energy, with the filtered noise used in the constraints for the SDC estimator. One problem with this approach is that the perceptual filter requires estimating an all pole model of the clean speech spectrum from the noisy signal.

2. FIXED GAIN CHARACTERISTICS

In this paper, we analyse linear filters calculated over a frame and applied to a noisy speech vector, $\mathbf{z}$, within that frame (see [8]). We assume that the noise, $\mathbf{w}$, is additive and uncorrelated with the speech, $\mathbf{y}$. We also assume that the noisy signal covariance matrix eigenvectors approximately diagonalize the clean signal and noise covariance matrices, $\mathbf{R}_y$ and $\mathbf{R}_w$ and that this is valid for the estimated matrices.

$$\mathbf{R}_z = \mathbf{R}_y + \mathbf{R}_w = \mathbf{V}\mathbf{\Lambda}_z\mathbf{V}^T = \mathbf{V}\left(\tilde{\mathbf{\Lambda}}_y + \tilde{\mathbf{\Lambda}}_w\right)\mathbf{V}^T \quad (1)$$

$$\begin{aligned} \tilde{\mathbf{\Lambda}}_y &= \mathbf{V}^T\mathbf{R}_y\mathbf{V} \approx \mathrm{diag}\{\tilde{\lambda}_{y_i}\}, \\ \tilde{\mathbf{\Lambda}}_w &= \mathbf{V}^T\mathbf{R}_w\mathbf{V} \approx \mathrm{diag}\{\tilde{\lambda}_{w_i}\} \end{aligned} \quad (2)$$

We denote as a *speech-dominated eigenvector* one along which the speech energy exceeds the noise energy as opposed to a *noise-dominated eigenvector*. Enhancement is achieved by

$$\hat{\mathbf{y}} = \mathbf{V}\mathbf{G}\mathbf{V}^T\mathbf{z} \quad \mathbf{G} = \mathrm{diag}\{g_i\} \quad g_i = \mathrm{f}(\tilde{\lambda}_{y_i}, \tilde{\lambda}_{w_i}) \quad (3)$$

where $\mathrm{f}(.)$ denotes some function of the eigenspectral SNR, the ratio of the clean speech to noise energy along the i^{th}

This project was funded through an ORS award.

eigenvector, $\tilde{\lambda}_{y_i}/\tilde{\lambda}_{w_i}$. As an example, the TDC gain function, $g^{(T)}$, and SDC gain function, $g^{(S)}$, are respectively given by

$$g^{(T)} = \frac{\tilde{\lambda}_y}{\tilde{\lambda}_y + \mu\tilde{\lambda}_w}, \quad \mu \geq 0 \tag{4}$$

$$g^{(S)} = \left(\frac{\tilde{\lambda}_y}{\tilde{\lambda}_y + \tilde{\lambda}_w}\right)^{\gamma/2}, \quad \gamma \geq 1 \tag{5}$$

where μ and γ control the trade-off between residual noise and signal distortion energy; higher values of the parameters provide increased noise suppression but with higher distortion. A sharper transition from high gain values to low ones is achieved with the following gain function [2]

$$g^{(E)} = \exp\left(-\nu\tilde{\lambda}_w/\tilde{\lambda}_y\right) \tag{6}$$

Like the parameter γ, ν controls the balance between distortion and residual noise. In all these gain functions, an empirically-determined fixed gain characteristic (i.e. μ, γ and ν are fixed) is applied.

Hu and Loizou [9][10] propose a variation of $g^{(T)}$ in which μ is a function of the segmental SNR, S dB, calculated over each frame. They vary μ in (5) so that it is higher for frames with low segmental SNR leading to higher suppression of background noise. The rule for choosing μ is given by

$$\mu = \begin{cases} \mu_0 - S/s_0 & -5 < S < 20 \\ 1 & S \geq 20 \\ 5 & S \leq -5 \end{cases} \tag{7}$$

where μ_0 and s_0 are experimentally determined. They show that their approach is similar to the noise power oversubtraction proposed in [1] where the oversubtraction factor is made to depend on segmental SNR. Although this leads to a family of gain functions, the parameterization of the gain function is still ad hoc.

3. ANALYSIS OF PERCEPTUAL-BASED GAIN CHARACTERISTICS

3.1. Using a psychoacoustic model

Several authors have suggested using the properties of human hearing to guide the selection of a gain function. Jabloun and Champagne [11] [8] choose to work with a psychoacoustic model to obtain the *masking threshold* for each frame, the energy level in the power spectral domain (PSD) below which any sound cannot be perceived by the hearing system due to masking by strong speech energy in that frame. The masking threshold is estimated in the PSD by applying a psychoacoustic model of hearing (ISO MPEG-1) to the clean speech spectrum or its estimate. This involves converting the frequency scale to the non-linear Bark scale to match the non-linear resolution of the auditory system and convolving with a spreading function to account for interband masking.

The authors propose a Frequency-to-Eigendomain transform (FET) to convert the mask from the PSD to the eigendomain. Having obtained the masking threshold for an eigenvector, $\tilde{\lambda}_m$, they modify (6) to use $\tilde{\lambda}_m$ instead of the clean speech energy $\tilde{\lambda}_y$, thus attenuating the noise only if it exceeds the masking threshold. They acknowldege that estimation errors are likely to arise for weak energy frames, like unvoiced fricatives, possibly leading to $\tilde{\lambda}_m$ exceeding $\tilde{\lambda}_y$. For this reason they take the minimum of the two in their expression for gain.

$$g^{(P)} = \exp\left(-\nu\frac{\tilde{\lambda}_w}{\min(\tilde{\lambda}_m, \tilde{\lambda}_y)}\right) \tag{8}$$

$$\begin{array}{lccl} \text{Now} & \min(\tilde{\lambda}_m, \tilde{\lambda}_y) & \leq & \tilde{\lambda}_y \\ \Rightarrow & \frac{\tilde{\lambda}_w}{\min(\tilde{\lambda}_m, \tilde{\lambda}_y)} & \geq & \frac{\tilde{\lambda}_w}{\tilde{\lambda}_y} \\ \Rightarrow & g^{(P)} & \leq & g^{(E)} \text{ from (6) and (8)} \end{array} \tag{9}$$

We show in (9) that the gain may be lower when the masking threshold is used. Thus for speech-dominated eigenvectors, the gain introduces slightly more attenuation than previously. Also, as noted above, λ_m is poorly estimated for noise-dominated eigenvectors and if λ_m exceeds $\tilde{\lambda}_y$, $g^{(P)}$ and $g^{(E)}$ would be identical. As a result, the noise suppression characteristics would not be better.

3.2. Problem with masking threshold in SDC derivation

To assess the weakness above, we adapt, to the eigenspectral domain, Hu and Loizou's [12] frequency-based technique of attenuating the noise to the level of the masking threshold. For each eigenvector, the gain g is calculated so that the residual noise along an eigenvector, $g^2\tilde{\lambda}_w$, does not exceed the masking threshold.

$$g^2\tilde{\lambda}_w \leq \tilde{\lambda}_m \Rightarrow g \leq \sqrt{\frac{\tilde{\lambda}_m}{\tilde{\lambda}_w}} \tag{10}$$

If $\tilde{\lambda}_m \geq \tilde{\lambda}_w$ then the speech already masks the noise so that a gain of 1 is used to avoid any distortion. If $\tilde{\lambda}_m < \tilde{\lambda}_w$, attenuation is needed and $g = \sqrt{\tilde{\lambda}_m/\tilde{\lambda}_w}$ from (10). If we choose the minimum of the masking threshold and clean speech energy for the same reasons as in (8) we can express the gain as

$$g^{(M)} = \min\left(1, \sqrt{\frac{\min(\tilde{\lambda}_m, \tilde{\lambda}_y)}{\tilde{\lambda}_w}}\right) \tag{11}$$

We have implemented (11) using the FET transform on the masking threshold in the PSD [8]. Listening tests reveal that for white noise at 5dB the residual noise is very audible, indeed more so than with a Wiener gain (SDC with γ=2 in (5), i.e. $g^{(S)}_{\gamma=2}$).

As reported in [8], for negative values of eigenspectral SNR, the masking threshold may be quite close to the speech

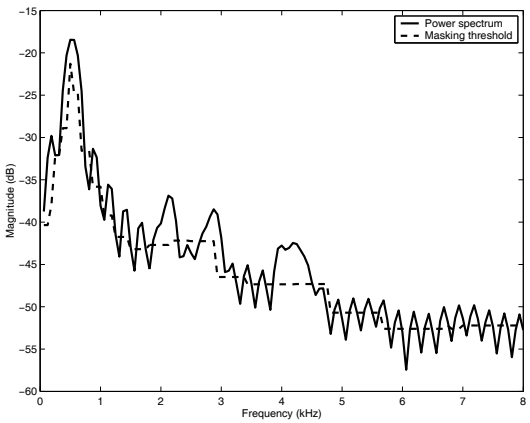

Fig. 1. *Power spectrum of clean speech (solid line) and masking threshold (dashed line) for a strong speech frame.*

energy. This can be seen, in the frequency domain, for the higher frequencies above 5kHz in Figure 1 where the masking threshold is shown as a dashed line. As a result, the following approximation can be derived for such eigenvectors

$$\sqrt{\frac{\tilde{\lambda}_m}{\tilde{\lambda}_w}} \approx \sqrt{\frac{\tilde{\lambda}_y}{\tilde{\lambda}_w}} \approx \left(\frac{\frac{\tilde{\lambda}_y}{\tilde{\lambda}_w}}{1 + \frac{\tilde{\lambda}_y}{\tilde{\lambda}_w}} \right)^{1/2} = g_{\gamma=1}^{(S)} \qquad (12)$$

The validity of this approximation is illustrated by plotting $g^{(M)}$ together with the curve for $g_{\gamma=1}^{(S)}$ in Figure 2. The masking threshold-based gain values are calculated using (11) for all eigenvectors of all frames for a speech extract corrupted by white noise. The values approach the $g_{\gamma=1}^{(S)}$ curve as eigenspectral SNR decreases, as we predicted. This particular curve gives the highest residual noise for this family of curves which explains why the noise level is quite high when using the gain in (11). The calculation of gain value based on the masking threshold is unsuitable because the gain expression (11) does not achieve sufficient noise suppression for negative values of eigenspectral SNR.

4. VARIABLE GAIN FUNCTION SELECTION

4.1. Expected energy-based approach

In this section, we develop a method for obtaining variable gain characteristics based on the masking threshold. We choose γ for each frame in (5) so that the energy in the frame is equal to an unbiased estimate of the clean speech energy, calculated for example by power spectral subtraction. We derive an expression for the expected value of the energy in a frame where

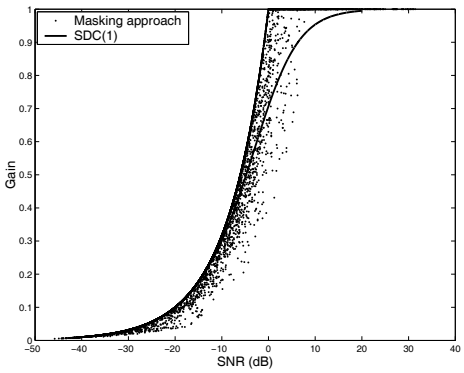

Fig. 2. *Plot of $g^{(M)}$ and $g_{\gamma=1}^{(S)}$ v/s eigenspectral SNR*

the gain used is as in (5).

$$\begin{aligned} \mathrm{trace}(E(\hat{\mathbf{y}}\hat{\mathbf{y}}^T)) &= \mathrm{trace}(\mathbf{V}_z\mathbf{G}\mathbf{V}_z^T E(\mathbf{z}\mathbf{z}^T)\mathbf{V}_z\mathbf{G}\mathbf{V}_z^T) \\ &= \mathrm{trace}(\mathbf{V}_z\mathbf{G}\mathbf{V}_z^T\mathbf{R}_z\mathbf{V}_z\mathbf{G}\mathbf{V}_z^T) \\ &= \mathrm{trace}(\mathbf{V}_z\mathbf{G}\mathbf{\Lambda}_z\mathbf{G}\mathbf{V}_z^T) \\ &= \mathrm{trace}(\mathbf{G}\mathbf{\Lambda}_z\mathbf{G}) \\ &= \mathrm{trace}(\mathbf{G}^2\mathbf{\Lambda}_z) \\ &= \sum \left(\frac{\tilde{\lambda}_y}{\tilde{\lambda}_y+\tilde{\lambda}_w}\right)^{\gamma} \lambda_z \end{aligned} \qquad (13)$$

We vary γ from the lowest value of 1 and select the value, γ^*, for which (13) is closest to the unbiased estimate of the clean speech energy. We place an upper limit of 4 on the value of γ^* to limit suppression, i.e. $1 \leq \gamma \leq 4$. In addition, we ensure that no suppression is applied if the masking threshold exceeds the noise energy level. We obtain $\tilde{\lambda}_m$ as in section 3.2 by calculating the masking threshold from the enhanced speech obtained using γ^* in (5). We decide on a gain value of $g_{\gamma^*}^{(S)}$ or 1 depending on whether our estimate for the noise energy, $\tilde{\lambda}_w$, exceeds the masking threshold, $\tilde{\lambda}_m$, or not.

In choosing γ^* as in the previous paragraph, we keep distortion for speech-dominated eigenvectors low since attenuation is not always applied. In addition, we improve on the masking threshold-based approaches (section 3) for noise suppression for noise-dominated eigenvectors since γ^* is likely to be much greater than 1 to decrease (13) to a desired level.

4.2. Results

In this section, we compare the performance of the three following schemes: the variable gain defined in (7), the masking threshold approach described in section 3.2, and the new technique proposed in section 4.1. Since the balance between speech distortion energy and residual noise energy is achieved differently in each case, we use these two energy metrics for comparison. A speech test file is corrupted with white noise for an input SNR of 5dB; the eigenspectral SNR values are calculated for each eigenvector of each frame. For each of the three schemes, we calculate the gain values from the

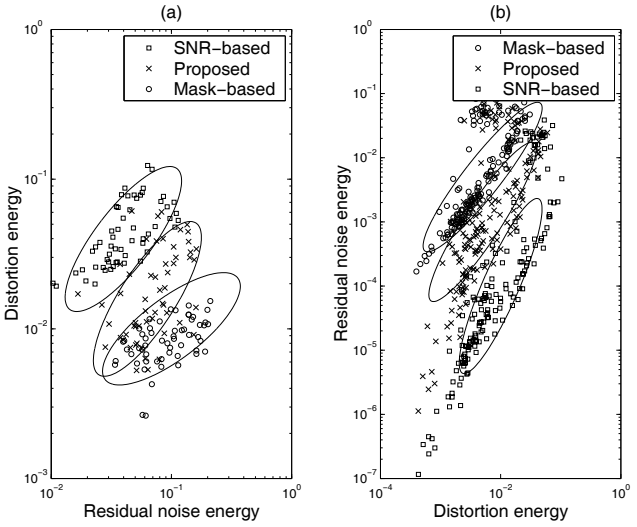

Fig. 3. Noise energy metrics for 3 different techniques for frames with (a) high values of eigenspectral SNR, i.e. *speech-dominated frames* (b) low values of eigenspectral SNR, i.e. *noise-dominated frames*

eigenspectral SNR values and using any other information required, for example the segmental SNR values for the variable gain approach (7), and the masking threshold curves for the other two schemes (sections 3.2 and 4.1). With these gain values, the distortion and residual noise energy expressions are calculated for each frame for the three cases and are plotted on the same scatter plots in Figure 3(a) for *speech-dominated frames* and Figure 3(b) for *noise-dominated frames*. Note that the axes are interchanged to have the appropriate metric on the y-axis, e.g. speech distortion energy for speech-dominated frames.

The desired result is to have low values of distortion energy for the speech-dominated frames and low values of residual noise energy for the noise-dominated frames. In the former case (Figure 3(a)), the cluster of points for the masking threshold approach is the lowest followed by that for the proposed technique and that for the variable gain approach, i.e. performance decreases in that order. This is to be expected since the first and second approaches base their decision to attenuate or not on the masking provided by the speech energy which is high for the speech-dominated frames. For the noise-dominated frames (Figure 3(b)), the performance order is reversed since the cluster of residual noise energy points is highest for the masking threshold-based approach. This stems from the inherent weakness of using the masking threshold in gain calculation for frames with low or no speech energy (section 3.2). The proposed approach gives distortion and residual noise values that lie between the two extremes and gives a good trade-off: the distortion energy for speech-dominated frames is kept low by exploiting the perceived noise level while the residual noise level in weak speech and noise-dominated frames is managed by trying to match a frame-wide desired energy level.

5. CONCLUSION

In this paper, we indicate that previously proposed gain functions for eigenspectral speech enhancement use a fixed parameter value that does not take the local signal information in consideration. Some approaches use the masking threshold in the gain function definition but we point out that this may allow an excess of residual noise while unnecessarily adding distortion energy. We propose an approach where the masking threshold is not used directly in the gain function calculation but where the parameter value is chosen so that the actual value of the enhanced speech energy is made to match an unbiased estimate of the clean speech energy. From the noise energy metrics, we conclude that our approach provides a good balance based on the speech or noise energy content along each eigenvector.

6. REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1979, vol. 4, pp. 208–211.

[2] Y. Ephraim and H.L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[3] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 159–167, 2000.

[4] A. Rezayee and S. Gazor, "An Adaptive KLT Approach for Speech Enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 87–95, 2001.

[5] Y. Hu and P.C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," in *Proceedings International Conference on Acoustics, Speech and Signal Processing*, 2002, vol. 1, pp. 573–576.

[6] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Processing Letters*, vol. 10, no. 4, pp. 104–106, 2003.

[7] Y. Hu and P.C. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 457–465, 2003.

[8] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 700–708, 2003.

[9] Y. Hu and P.C. Loizou, "Perceptually motivated subspace approach for speech enhancement," in *Proceedings International Conference on Spoken Language Processing*, 2002, pp. 1797–1800, ICSLP.

[10] Y. Hu and P.C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[11] F. Jabloun and B. Champagne, "A perceptual signal subspace approach for speech enhancement in colored noise," in *International Conference on Acoustics, Speech and Signal Processing*, 2002, vol. 1, pp. 569–572.

[12] Y. Hu and C. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 270–273, 2004.