Residual Noise Suppression Using Psychoacoustic Criteria

Tim Haulick, Klaus Linhard and Peter Schrögmeier

Daimler Benz AG, Research and Technology, Wilhelm-Runge-Str. 11
D-89081 Ulm, Germany
e-mail: haulick@dbag.ulm.daimlerbenz.com

ABSTRACT

Speech enhancement techniques using spectral subtraction have the drawback of generating residual noise with a musical character, so-called musical noise. We developed a new post-processing method for suppressing this musical residual noise. In this method, the auditory masking threshold is calculated twice, once before the spectral subtraction and once again afterwards. This ensures that all audible spectral signal components above the thresholds are detected. Audible components which are only present at the output are candidates for musical noise. Depending on their spectral bandwidth and time duration, they may be processed additionally. Using this post-processing, the distortion of the speech signal is not noticeable and musical noise is not audible even at low signal-to-noise ratios of about 0 dB.

1. INTRODUCTION

Additive wideband acoustic noise is a common problem in modern hands-free telecommunication and automatic speech recognition. Acceptable speech quality can only be attained through noise reduction. Spectral subtraction is especially attractive due to its simplicity.

We refer to spectral magnitude subtraction, that is the subtraction of the estimated noise magnitude from the noisy speech magnitude [1]. This subtraction is usually implemented as a frequency domain multiplication of the noisy speech signal Y with real filter coefficients H.

If we denote the Fourier transformed noisy speech signal as the sum of the original speech signal and noise

$$Y_{k,i} = S_{k,i} + N_{k,i}$$

then the filter coefficients are computed as

$$H_{k,i} = 1 - \sqrt{\frac{E[|N_i|^2]}{|Y_{k,i}|^2}} = 1 - \frac{\bar{N}_i}{|Y_{k,i}|}$$

with k and i denoting the discrete time segment index and the discrete frequency index, respectively. $\bar{N}_i$ is usually adapted by recursive time averaging in speech pauses. The multiplication is equivalent to the subtraction

$$\hat{S}_{k,i} = Y_{k,i} - \bar{N}_i \frac{Y_{k,i}}{|Y_{k,i}|} = (|Y_{k,i}| - \bar{N}_i) \cdot \phi(Y_{k,i})$$

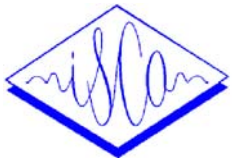
In general, the actually present noise $N_{i,k}$ and the average noise $\bar{N}_i$ are different, and $N_{k,i}$ is different from $N_{k,i\pm j}$; $(j = 1, 2, \ldots)$. Thus, the spectral differences, $N_{k,i} - \bar{N}_i$, appear and disappear with the block processing rate of typically 100 Hz if we process signal segments every 10 msec. These differences have some dominant peaks with limited spectral bandwidth and thus sound like tones which are switched on and off (musical noise).

Standard spectral subtraction uses noise overestimation and noise masking to reduce musical tones. In noise overestimation, aN rather than just N is subtracted with $1 < a \leq 4$. However, larger values of a yield higher speech distortion, [2], [3]. Noise masking is accomplished using a lower bound b on the coefficient H with $0 < b \leq .25$, but b limits the maximum amount of noise reduction. Standard noise overestimation and noise masking are thus not sufficient to achieve high noise reduction, eliminate musical tones, and at the same time yield good speech quality.

A first psychoacoustically based approach for spectral subtraction is reported in [7]. The loudness of the estimated noise is subtracted from the loudness of the noisy speech signal. Other approaches, suggested in [8] and [9], use the auditory masking threshold to avoid suppression of spectral components below this threshold, and thus suppress only audible noise and minimize signal distortion. We propose a method for detecting any new audible spectral components at the output after signal processing, e.g. spectral subtraction. The auditory masking threshold is calculated at both the input and output signals. Audible spectral components lie above this threshold. New output components are those audible only at the output. These new components are candidates for musical residual noise and may be postprocessed. This approach has the potential of being independent of the particular method of noise reduction. In the case of spectral subtraction and musical noise, we use a bandwidth and time criterion to decide if the new components are new audible speech components or more likely musical noise which needs to be suppressed.

2. MASKING MODEL

A masking model for transform coding of audio signals is described in detail in [6]. A perceptual threshold is used to minimize the bit rate of the coder, thus



EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

allowing quantization noise only below this perceptual threshold. There are four steps involved in calculating this threshold:

- The frequency power spectrum is converted into the bark-related critical band spectrum by adding up the energies of all frequency components within each band.
- Masking effects across critical bands are approximated by convoluting the critical band spectrum with the so called spreading function.
- An offset factor is applied to the masking threshold to take into account the different masking properties of tones and noise.
- The final theshold is determined by renormalizing the results and comparing them to the absolute hearing threshold.

In order to reduce residual musical noise, we need an audible threshold for musical residual noise for masking. The noisy speech plays the role of the masking signal. A practical example of this is speech corrupted by road noise. For this case, the masking threshold has been calculated by [6] (see Fig. 1). Auditory experiments have been conducted using this road noise as a masker and several sinusoidal test stimuli of different frequencies. The calculated threshold is close to the average threshold of several subjects indicated with '*' in Fig. 1. If the sinusoidal stimuli are exchanged with short sinusoidal tones of approximately 10 ms duration, the threshold is approximately 10 dB higher (see 'o' in Fig. 1). The short sinusoidal tones are a simplified model of musical noise. Thus adding 10 dB to the caluculated threshold yields a good first estimation of the audible threshold for musical noise.

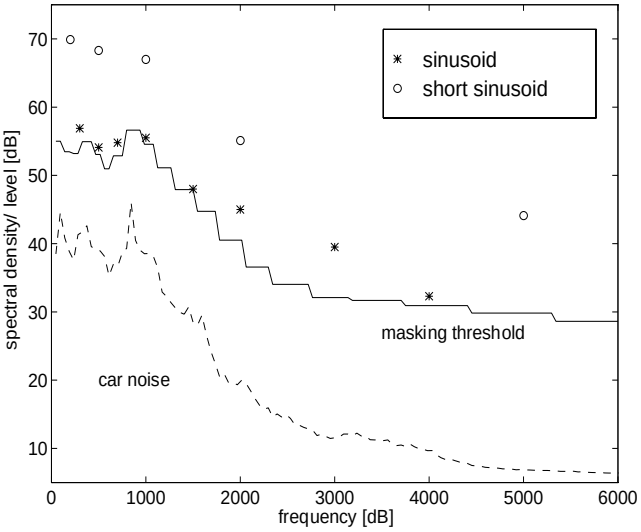


Figure 1. *Masking threshold, car noise masker and sinusoid test signal*

In calculating the masking threshold, [6] distinguishes between two situations: "tone masking noise" and "noise masking a tone." Depending on the situation, a corresponding offset factor is applied to the threshold. To avoid practical problems with the detection of the situation, it is appropriate to use only the more conservative threshold connected to "noise masking tone." This is the case during speech pauses where musical noise is especially annoying.

3. POST-PROCESSING

After calculating the masking threshold at both the input and the output, all audible spectral components are detected. Fig. 2 shows an example of a noisy speech segment. Components above the threshold in both figures should be speech components, denoted by 's'. Components only present at the output are candidates for musical noise, denoted by 'm'. In order to avoid accidentally identifying a new speech component at the output as musical noise and thus suppressing it, a short-time stationarity and bandwidth criterion is applied. In order to be classified as musical, noise must have a bandwidth of not more than approximately 300 Hz. Otherwise, it is more likely that this "noise" is unvoiced speech sound. If succeeding filter coefficients fulfill the condition

$$H_{k-1,i} \wedge H_{k,i} \geq .55$$

the corresponding spectral component is assumed to be speech and is not suppressed. If a new spectral component at the output meets the bandwidth condition and the single spectral lines are only present once in succession, its values are set to the spectral floor:

$$\hat{S}_{k,i} = b\, Y_{k,i}; \quad (b = .25)\ .$$

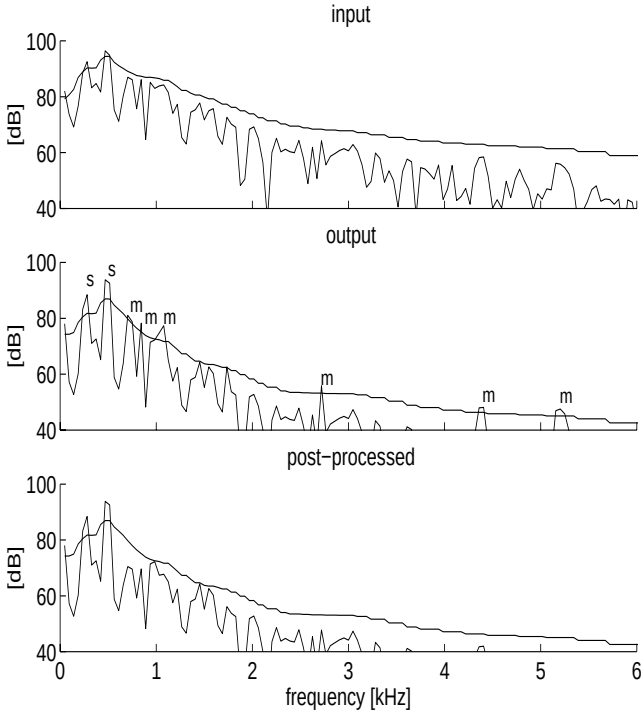


Figure 2. *Short term spectral densities of noisy speech at input, output and after post-processing*

4. RESULTS

Intensive subjective listening tests have been conducted to evaluate optimal parameters, e.g. bandwidth and stationarity criteria, offset factor, spectral resolution, etc. The tests were performed using a number of noisy speech sentences (S1-S4) as described in Table 1, generated by adding different car noises to a clean speech signal. After optimization, several tests were performed with fixed parameters. Using objective measures and informal listening tests the results of the psychoacoustically based post processing approach (PSY) have been compared to the results of three other enhancement schemes: standard spectral subtraction (SPS), an optimized spectral subtraction (Boll) as described in [1] and the estimator proposed by Ephraim and Malah [4],[5] (MMSE). Two objective measures were calculated – the global signal-to-noise ratio (SNR) which gives the overall noise reduction, and the segmental SNR ($\mathrm{SNR}_{\mathrm{seg}}$) as a performance measure with respect to signal distortion. With $s(k)$ and $\hat{s}(k)$ denoting the clean speech signal and the processed output, respectively, the segmental SNR is defined by

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{10}{M}\sum_{m=0}^{M-1} \lg \frac{\sum_{k=0}^{K-1} s(k+Km)^2}{\sum_{k=0}^{K-1} (s(k+Km)-\hat{s}(k+Km))^2},$$

where K (=256) is the number of signal samples in a frame and M is the number of frames.

As shown in Table 2 and Table 3, both the optimized spectral subtraction (Boll) as well as the proposed psychoacoustically based approach (PSY) provide a significantly reduced residual noise with only a slightly degraded segmental SNR compared to standard spectral subtraction. For all signals the MMSE estimator indicates a superior output SNR but also the worst segmental SNR. This was also confirmed by informal listening tests which showed a very pleasant residual noise but also an annoying reverberant character of the processed speech signal in case of the MMSE estimator. The output signal of the proposed post-processing approach (PSY) was found similar to the MMSE estimator without noticeable musical artifacts, whereas with the optimized spectral subtraction approach (Boll) musical noise was still audible.

The performance of the proposed enhancement method versus standard spectral subtraction is illustrated in Fig. 3. Figure 3(a) shows a time-frequency decomposition of 14400 samples taken from a noisy speech signal (S3). Figure 3(b) shows the corresponding spectral plot of the enhanced speech segment using standard spectral subtraction. Although a substantial amount of noise has been removed, some dominant spectral peaks remain in the output signal thus becoming audible as musical noise. The enhanced signal after additionally post-processing is shown in Figure 3(c). As it is clearly visible, the proposed algorithm has significantly reduced the residual noise without noticeable impairment of the speech signal.

ID	*driving condtion*			*SNR*
	ENGINE	FAN	SUNROOF	(dB)
S1	off	high	closed	10.5
S2	100 km/h	off	closed	9.0
S3	100 km/h	off	open	5.5
S4	160 km/h	off	closed	0.5

Table 1. *Description of the degraded speech signals used in the test*

ID	*Global SNR*				
	Unproc.	SPS	Boll	PSY	MMSE
S1	10.5	19.2	21.3	21.5	23.2
S2	9.0	17.5	19.8	20.1	21.8
S3	5.5	12.6	15.1	15.6	17.3
S4	0.5	7.1	9.4	9.8	11.2

Table 2. *Global SNR (in dB) with respect to the unprocessed input for 4 different enhancement schemes ($b = .25$)*

ID	*Segmental SNR*				
	Unproc.	SPS	Boll	PSY	MMSE
S1	10.2	12.4	12.2	12.2	10.9
S2	9.7	12.3	12.0	12.0	11.0
S3	6.7	9.2	9.0	9.1	8.3
S4	3.3	7.2	7.1	7.1	6.7

Table 3. *Segmental SNR (in dB) with respect to the unprocessed input for 4 different enhancement schemes ($b = .25$)*

5. CONCLUSIONS

In this paper, a new post-processing method for suppressing musical artifacts after signal processing, e.g. spectral subtraction has been presented. The method utilizes the auditory masking threshold of the human ear to detect new audible spectral components in the output signal. In combination with a simple time-/bandwith criterion, the proposed algorithm destinguishes reliably between musical noise and the useful signal, thus allowing to suppress effectively musical artifacts without noticeably affecting the speech signal even at low signal-to-noise ratios of about 0 dB.

REFERENCES

[1] Boll, S. F.: *A Spectral Subtraction Algorithm for Suppression of Acoustic Noise in Speech*, Proc. on ICASSP 1979, pp. 200–203

[2] Boll, S. F.: *Suppression of Acoustic Noise in Speech Using Spectral Subtraction*, IEEE Trans. on ASSP, Vol. 27, No. 2, pp. 113–120, April 1979

[3] Berouti, M.; Schwartz, R.; Makhoul, J.: *Enhancement of Speech Corrupted by Acoustic Noise*, IEEE ICASSP 1979, pp. 208–211

[4] Ephraim, Y.; Malah, D.: *Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator*, IEEE Trans. on ASSP, Vol. 32, No. 6, pp. 1109-1121, Dec. 1984

[5] Cappé, O.: *Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor*, IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 345-349, Apr. 1994

[6] Johnston, J.: *Transform Coding of Audio Signals Using Perceptual Noise Crtieria*, IEEE J. on Select. Areas in Commun., Vol. 6, pp. 314–323, Feb. 1988

[7] Petersen, T.; Boll, S. F.: *Acoustic Noise Suppression in the Context of a Perceptual Model*, IEEE ICASSP 1981, pp. 1086–1088

[8] Tsoukalas, D.; Paraskevas, M.; Mourjopoulos, J.: *Speech Enhancement using Psychoacoustic Criteria*, IEEE ICASSP 1993, pp. II 359–II 362

[9] Azirani, A. A.; Le Bouquin Jeannes, R.; Faucon, G.: *Optimizing Speech Enhancement by Exploiting Masking Properties of the Human Ear*, IEEE ICASSP 1995, pp. 800–803

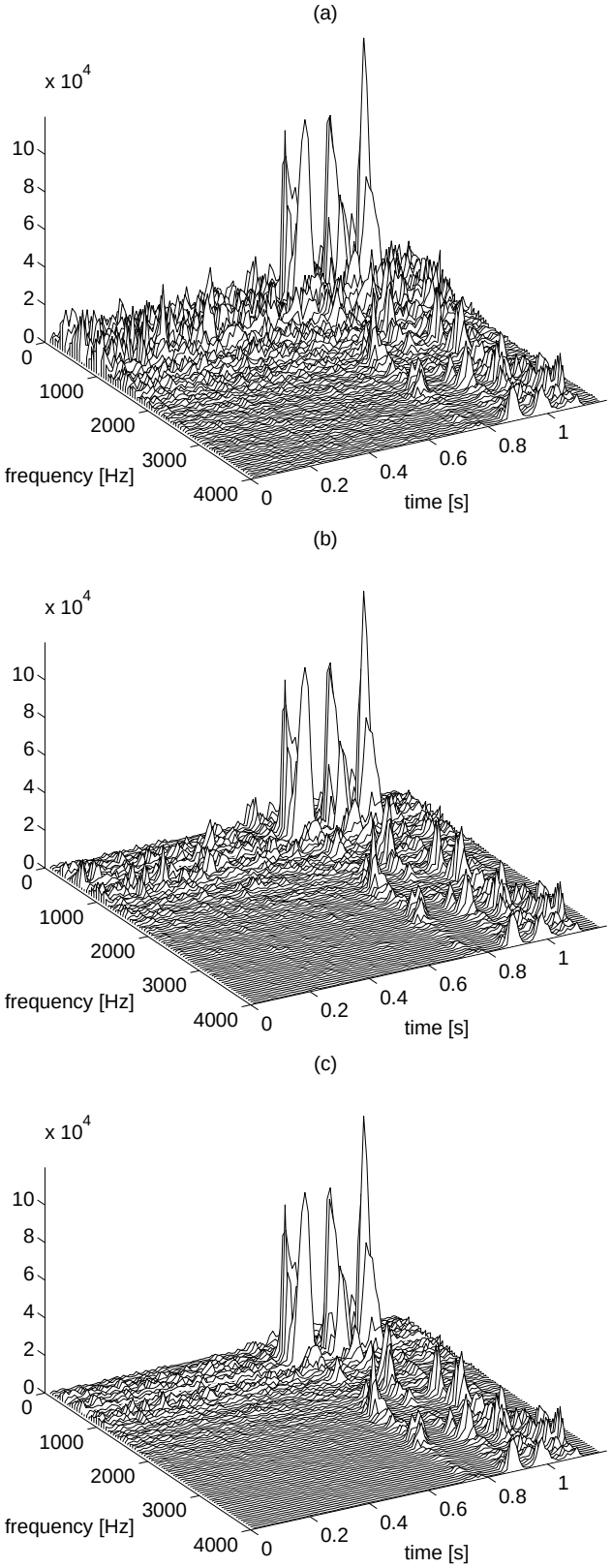


Figure 3. *Time-frequency decomposition of 1.2s of speech. (a) Noisy speech signal; (b) enhanced speech signal using standard SPS; (c) enhanced speech signal after post-processing*