WIDEBAND NOISE SUPPRESSION SUPPORTED BY ARTIFICIAL BANDWIDTH EXTENSION TECHNIQUES

Thomas Esch, Florian Heese, Bernd Geiser, Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{esch|heese|geiser|vary}@ind.rwth-aachen.de

ABSTRACT

This contribution presents a wideband (50 Hz – 7 kHz) speech enhancement system that is operating in the frequency domain. As a novel feature, techniques known from artificial bandwidth extension (BWE) are used to improve the spectral estimation process by exploiting the statistical dependencies between the low band (50 Hz – 4 kHz) and the high band (4 – 7 kHz). Conventional noise suppression is used in the low band, while a novel approach is applied to the high band. Features from the processed (enhanced) low band signal are extracted and used to estimate subband energies of the high band. The weighting gains determined from these energy estimates are adaptively combined with conventional gains obtained in addition for the high band. The performance of the proposed method is shown to be consistently better than the conventional approach, especially at low input SNR values.

Index Terms— Wideband speech enhancement, noise reduction, artificial bandwidth extension

1. INTRODUCTION

The quality of today's telephone speech was designed to achieve a sufficient intelligibility. The acoustic bandwidth in telephony systems is typically limited to the frequency range between 300 Hz and 3.4 kHz. However, this typical "telephone sound" cannot satisfy the increased demands as the perceived speech quality is considerably reduced compared to the full audio bandwidth. As a reasonable compromise, various wideband (50 Hz – 7 kHz) speech codecs have been developed in the past (e.g., the Adaptive Multi-Rate (AMR) Wideband Codec) and are about to be introduced in current mobile networks. Nevertheless, most of these codecs are mainly designed for nearly noise-free input speech signals and do not perform well when the input signal is degraded by acoustic background noise. In order to improve the listening comfort and to keep the high quality also in noisy environments, noise suppression techniques are required for wideband communication systems.

One of the popular methods for enhancing degraded speech is based on modeling the noisy input coefficients in the short-time Fourier transform (STFT) domain and to apply individual adaptive gains for each frequency bin. Most of the rules proposed in literature have been derived for low band (50 Hz – 4 kHz) signals under certain assumptions about the statistics of the speech and noise signals, e.g., [1–3]. When it comes to wideband noise reduction, an established method is to double sampling rate and transform length and to apply the low band algorithms also for higher frequencies. Thereby, neither the unequal spectral energy distribution of a speech and noise signal nor the properties of the human auditory system are considered. For typical realistic noise sources, it can be shown that the signal-to-noise ratio (SNR) significantly degrades beyond 4 kHz leading to imprecise noise reduction and fluctuating weighting gains that result in the increased occurrence of *musical noise* especially at higher frequencies. So far, only a very limited number of proposals are known which take into account the afore mentioned aspects when enhancing wideband speech signals, e.g., [4].

It is known from the field of speech coding that the spectral dependencies of speech signals can be exploited to recover missing high frequency components by utilizing only the low band speech signal. This technique, called artificial bandwidth extension (BWE), aims at increasing the perceived speech quality if only the low band signal is available. In this paper, wideband speech enhancement is combined with techniques known from artificial BWE. While a conventional noise suppression technique is used in the low band, a joint approach is applied for the speech enhancement in the high band (4 – 7 kHz). Based on a trained hidden Markov model (HMM), features from the processed (enhanced) low band signal are extracted and used to estimate subband energies of the high band speech signal. The resulting weighting gains determined from these energy estimates are adaptively combined with conventional gains for the high band. The remainder of this paper is organized as follows: In Sec. 2, a brief overview of the proposed system is given. Section 3 comprises the procedure of the combined noise suppression in the high band in detail. Experimental results are shown in Sec. 4 and conclusions are drawn in Sec. 5.

2. SYSTEM OVERVIEW

A simplified block diagram of the proposed wideband speech enhancement system is depicted in Fig. 1. It is assumed that the noisy input signal $y(k)$ consists of the clean speech signal $s(k)$ which is degraded by an additive noise signal $n(k)$ according to:

$$y(k) = s(k) + n(k), \tag{1}$$

where k is the discrete time index. Different processing schemes are applied in the low band (50 Hz – 4 kHz) and the high band (4 –

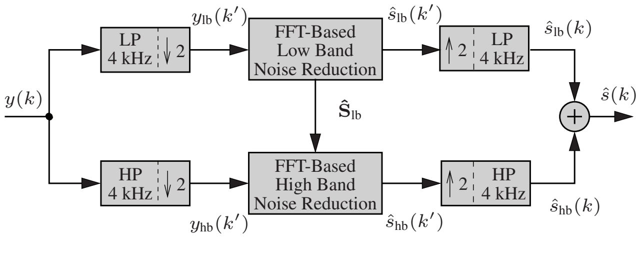
$y_{\text{lb}}(k')$
LP 4 kHz ↓2
FFT-Based Low Band Noise Reduction
$\hat{s}_{\text{lb}}(k')$
↑2 LP 4 kHz
$\hat{s}_{\text{lb}}(k)$
$\hat{s}(k)$
$y(k)$
$\hat{\mathbf{S}}_{\text{lb}}$
+
HP 4 kHz ↓2
FFT-Based High Band Noise Reduction
↑2 HP 4 kHz
$y_{\text{hb}}(k')$
$\hat{s}_{\text{hb}}(k')$
$\hat{s}_{\text{hb}}(k)$

Fig. 1. Wideband noise reduction using different techniques in low band and high band exploiting spectral dependencies.

This work was supported by Nokia, Tampere, Finland.

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

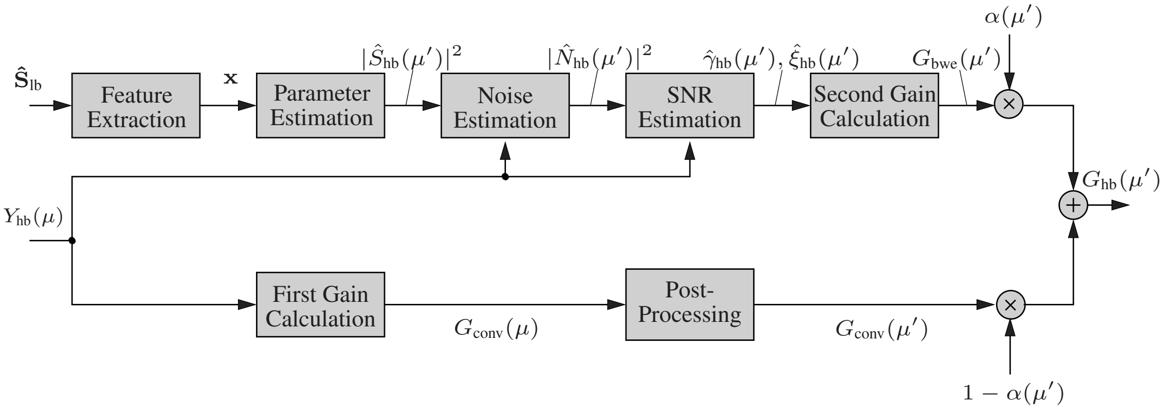

Fig. 2. High band noise reduction exploiting spectral dependencies between low band and high band.

7 kHz). Therefore, a 2-channel finite impulse response (FIR) quadrature mirror filter (QMF) bank with critical sampling and perfect reconstruction is used to split the wideband signal $y(k)$ into the low band and the high band signal. Due to the decomposition, individual analysis-synthesis structures and different algorithms can be used in each band enabling the re-use of existing low band noise reduction systems. After downsampling the lowpass and highpass filtered signals by a factor of 2, a conventional noise reduction technique is applied to the low band signal $y_{\text{lb}}(k')$ where k' represents the discrete time index in the subsampled domain. In the high band, the noisy signal $y_{\text{hb}}(k')$ is enhanced by using additional information from the improved low band signal. For this, features from the vector $\hat{\mathbf{S}}_{\text{lb}}$, consisting of the spectral coefficients from the processed low band signal, are extracted as will be explained in the next section.

In both bands, the noise suppression is performed in the frequency domain. Therefore, $y_{\text{xx}}(k')$ is segmented into overlapping frames of length L_{F}, where the index 'xx' denotes either the low band 'lb' or the high band 'hb'. After windowing and zero-padding, the fast Fourier transform (FFT) of length M_{F} is applied to these frames. Hence, the spectral coefficients of the noisy input signal at frequency bin μ and frame λ are given by:

$$Y_{\text{xx}}(\lambda,\mu) = S_{\text{xx}}(\lambda,\mu) + N_{\text{xx}}(\lambda,\mu), \qquad (2)$$

where $S_{\text{xx}}(\lambda,\mu)$ and $N_{\text{xx}}(\lambda,\mu)$ represent the spectral coefficients of the speech and the noise signal. For the sake of brevity, the frame index λ is omitted in the following.

The respective enhanced signals $\hat{s}_{\text{lb}}(k')$ and $\hat{s}_{\text{hb}}(k')$ are upsampled and lowpass and highpass filtered again. Finally, both signals are added in order to obtain the enhanced wideband signal $\hat{s}(k)$.

Noise Type	Average Deviation of the Low Band SNR from the High Band SNR for	
	Male Speakers	Female Speakers
Cockpit	+15.39 dB	+13.98 dB
Babble	+0.55 dB	-0.86 dB
Factory	+12.55 dB	+11.14 dB
Buccaneer	+15.64 dB	+14.23 dB
WGN	+26.81 dB	+25.39 dB

Table 1. SNR deviation of the low band from the high band for different noise types. For the measurement, six speech signals (three male and three female speakers) from the NTT database were used. The noise signals have been taken from the NOISEX-92 database.

3. JOINT NOISE REDUCTION IN THE HIGH BAND

The main energy of a speech signal is usually located in the frequency range between 500 Hz and 3 kHz. Assuming that the energy of speech signals declines stronger than the energy of noise signals beyond 3 kHz, the SNR in the low band is usually significantly higher than in the high band. Table 1 shows some quantitative examples of how much the SNR in the low band is better than in the high band for different speakers and different noise environments. It can be seen that in most cases the SNR significantly degrades in the high band which leads to an imprecise noise reduction and fluctuating weighting gains if solely a conventional noise suppression technique is applied to the higher frequencies. To counteract this problem, a joint noise reduction method is presented in this paper for the high band signal which makes use of the spectral dependencies between low band and high band.

Figure 2 shows the basic principle of the combined noise reduction scheme in the high band. The analysis and synthesis structure remains the same as for the low band signal. After the transformation into the frequency domain, two separate noise suppression methods are applied to the noisy high band spectrum $Y_{\text{hb}}(\lambda)$ resulting in the calculation of the high band weighting gains $G_{\text{hb}}(\mu')$ where μ' represents the subsampled frequency index as will be explained later.

As depicted in Fig. 2, a first (conventional) and a second (new) gain calculation is performed for the high band spectrum. The conventional noise reduction technique includes noise power estimation (e.g., [5]), SNR estimation (e.g., [2]) and the calculation of the weighting gains $G_{\text{conv}}(\mu)$ (e.g., [1–3]). In order to reduce the variance of the weighting gains, a post processing stage follows in which the frequency resolution is decreased from M_{F} to M'_{F}. Adjacent frequency-bins are combined using overlapping Hann windows of the same length. The reduction of the frequency resolution allows for an increased suppression of musical tones and corresponds to the properties of our human auditory system where the frequency selectivity decreases with higher frequencies.

In the upper branch of Fig. 2, artificial BWE techniques are used to perform the second gain calculation (see next section for details). All required processing steps are thereby performed at the reduced frequency resolution M'_{F} as well. The resulting weighting gains $G_{\text{bwe}}(\mu')$ are adaptively combined with $G_{\text{conv}}(\mu')$ according to:

$$G_{\text{hb}}(\mu') = \alpha(\mu') \cdot G_{\text{bwe}}(\mu') + \big(1-\alpha(\mu')\big) \cdot G_{\text{conv}}(\mu'), \qquad (3)$$

where $\alpha(\mu') \in [0,1]$ represents a cross-fading factor that is frame and frequency dependent as will be shown later. Finally, the frequency resolution of the high band weighting gains $G_{\text{hb}}(\mu')$ is expanded back from M'_F to the original resolution M_F using overlap-add of scaled Hann windows. A spectral weighting of the noisy high band coefficients $Y_\mathrm{hb}(\mu)$ with the weighting gains $G_\mathrm{hb}(\mu)$ yields an estimate $\hat{S}_\mathrm{hb}(\mu)$ of the clean high band coefficients $S_\mathrm{hb}(\mu)$:

$$\hat{S}_\mathrm{hb}(\mu) = Y_\mathrm{hb}(\mu) \cdot G_\mathrm{hb}(\mu). \tag{4}$$

An inverse fast Fourier transform (IFFT) and overlap-add is applied to obtain the enhanced signal $\hat{s}_\mathrm{hb}(k')$ in the time domain.

3.1. Noise Reduction Exploiting Spectral Dependencies

In order to exploit the dependencies in the frequency domain between low band and high band, techniques known from artificial BWE are applied for the wideband speech enhancement. The main principle that is used here for the BWE is partly included in [6]. The concept consists of estimating high band signal parameters based on features that are extracted only from the enhanced low band signal using a trained Hidden Markov Model (HMM).

Usually, representations of the spectral envelope of the low band signal are used as features that are extracted on a frame-by-frame basis [7]. In this realization, the feature vector $\mathbf{x}$ from the low band consists of N_C mel-frequency cepstral coefficients (MFCCs) and the zero-crossing rate (ZCR) of the low band signal. According to [6], a trained HMM is used to estimate the feature vector $\mathbf{y}$, representing the M'_F subband energies of the high band signal. Let $\mathbf{X} = \{\mathbf{x}(1), ..., \mathbf{x}(\lambda)\}$ be a sequence of feature vectors from the low band of frames 1 to λ. The criterion for MMSE estimation of a vector $\mathbf{y}$, with given observations $\mathbf{X}$ is $\mathcal{E}\{||\mathbf{y} - \hat{\mathbf{y}}||^2 | \mathbf{X}\} = \min$, where $\hat{\mathbf{y}}$ is the respective estimate. The solution to this optimization problem is the conditional expectation $\mathbf{y}_\mathrm{MMSE} = \mathcal{E}\{\mathbf{y}|\mathbf{X}\}$. Using a precomputed codebook $\mathcal{C} = \{\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_{M_\mathrm{C}}\}$ for the vectors $\mathbf{y}$ (e.g., obtained with the LBG algorithm [8]), this MMSE estimate can be expressed as [7, 9]:

$$\hat{\mathbf{y}}_\mathrm{MMSE} = \sum_{\hat{\mathbf{y}}_i \in \mathcal{C}} \hat{\mathbf{y}}_i \cdot P(\hat{\mathbf{y}}_i | \mathbf{X}), \tag{5}$$

which essentially is a weighted sum over the M_C centroids of the codebook $\mathcal{C}$. Thereby, the weights $P(\hat{\mathbf{y}}_i|\mathbf{X})$ specify a posteriori probabilities which can be calculated using HMM techniques [6].

Once the instantaneous energies of the M'_F subbands $\hat{\mathbf{y}} = \{|\hat{S}_\mathrm{hb}(0)|^2, ..., |\hat{S}_\mathrm{hb}(M'_\mathrm{F} - 1)|^2\}$ have been estimated, they are used to estimate the noise power in the high band signal:

$$|\hat{N}_\mathrm{hb}(\mu')|^2 = \max\left(|Y_\mathrm{hb}(\mu')|^2 - |\hat{S}_\mathrm{hb}(\mu')|^2, 0\right), \tag{6}$$

with $0 \leq \mu' \leq M'_\mathrm{F} - 1$. Finally, the *a posteriori* SNR $\gamma(\mu')$ and *a priori* SNR $\xi(\mu')$ can be estimated according to:

$$\hat{\gamma}_\mathrm{hb}(\mu') = \frac{|Y_\mathrm{hb}(\mu')|^2}{|\hat{N}_\mathrm{hb}(\mu')|^2} \quad \text{and} \quad \hat{\xi}_\mathrm{hb}(\mu') = \frac{|\hat{S}_\mathrm{hb}(\mu')|^2}{|\hat{N}_\mathrm{hb}(\mu')|^2}, \tag{7}$$

which are required in order to calculate the weighting gains $G_\mathrm{bwe}(\mu')$ to be used in Eq. 3.

3.2. Cross-Fading Factor

It has already been mentioned that the two weighting gains $G_\mathrm{conv}(\mu')$ and $G_\mathrm{bwe}(\mu')$ are adaptively combined using the cross-fading factor $\alpha(\mu')$, see Eq. 3. In the following, the ideal cross-fading factor $\alpha_\mathrm{opt}(\mu')$ is defined as:

$$\alpha_\mathrm{opt}(\mu') = \frac{(G_\mathrm{opt}(\mu') - G_\mathrm{conv}(\mu'))^2}{(G_\mathrm{opt}(\mu') - G_\mathrm{conv}(\mu'))^2 + (G_\mathrm{opt}(\mu') - G_\mathrm{bwe}(\mu'))^2}, \tag{8}$$

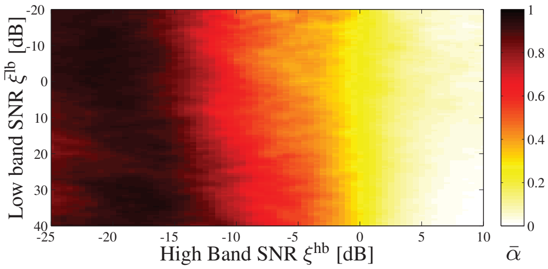

Fig. 3. Visualization example of look-up table to determine $\bar{\alpha}$.

where $G_\mathrm{opt}(\mu')$ represents the ideal weighting gain that could in theory (or by a dedicated simulation setup) be determined from the clean speech and noise signal according to the ideal *a posteriori* SNR $\gamma^\mathrm{hb}_\mathrm{opt}(\mu')$ and *a priori* SNR $\xi^\mathrm{hb}_\mathrm{opt}(\mu')$:

$$\gamma^\mathrm{hb}_\mathrm{opt}(\mu') = \frac{|Y_\mathrm{hb}(\mu')|^2}{|N_\mathrm{hb}(\mu')|^2} \quad \text{and} \quad \xi^\mathrm{hb}_\mathrm{opt}(\mu') = \frac{|S_\mathrm{hb}(\mu')|^2}{|N_\mathrm{hb}(\mu')|^2}, \tag{9}$$

which are also determined at the reduced frequency resolution M'_F by combining adjacent frequency bins as before. If the conventional noise suppression technique performs better than the BWE approach, i.e., $(G_\mathrm{opt} - G_\mathrm{conv})^2 < (G_\mathrm{opt} - G_\mathrm{bwe})^2$, α_opt in Eq. 8 tends to smaller values leading to a stronger weighting of G_conv in Eq. 3 and vice versa.

In order to estimate the optimal cross-fading factor in a realistic scenario, first $\alpha_\mathrm{opt}(\mu')$ is recorded in a training process for every frame λ and every subband μ' together with the respective subband SNR $\xi^\mathrm{hb}_\mathrm{opt}(\mu')$ of the high band and the averaged SNR $\bar{\xi}^\mathrm{lb}_\mathrm{opt}$ of the low band:

$$\bar{\xi}^\mathrm{lb}_\mathrm{opt} = \frac{1}{M_\mathrm{F}} \sum_{\mu=0}^{M_\mathrm{F}-1} \frac{|S_\mathrm{lb}(\mu)|^2}{|N_\mathrm{lb}(\mu)|^2}. \tag{10}$$

Based on the training data, a look-up table for the estimation of $\alpha(\mu')$ is generated for every subband. Therefore, $\xi^\mathrm{hb}_\mathrm{opt}(\mu')$ and $\bar{\xi}^\mathrm{lb}_\mathrm{opt}$ are quantized (e.g., 1 dB step size) and the associated values for $\alpha_\mathrm{opt}(\mu')$ are averaged within the quantization levels. At the end, the final look-up table provides one estimate $\bar{\alpha}(\mu')$ for each quantized combination of $\xi^\mathrm{hb}_\mathrm{opt}(\mu')$ and $\bar{\xi}^\mathrm{lb}_\mathrm{opt}$. A typical example of this two-dimensional look-up table can be seen in Fig. 3. The figure demonstrates a strong correlation between the averaged factor $\bar{\alpha}$ and the two SNR quantities showing that the BWE approach in Eq. 3 is preferred with a decreasing high band SNR. Moreover, in the high band SNR range -15 dB $\leq \xi^\mathrm{hb} \leq 0$ dB, it can be seen that the cross-fading factor $\bar{\alpha}$ becomes larger for higher low band SNR values $\bar{\xi}^\mathrm{lb}$ showing that the BWE (trained with clean speech) performs better the higher the input SNR is in the low band.

In a real application, $\xi^\mathrm{hb}_\mathrm{opt}$ and $\bar{\xi}^\mathrm{lb}_\mathrm{opt}$ are not available. Here, the respective SNR estimates of the conventional noise suppression techniques in the low band and high band are utilized to determine $\bar{\alpha}(\mu')$ using a pre-trained look-up table.

4. RESULTS

In principle, any noise reduction technique can be applied within the proposed system to perform the suppression in the low band and to estimate the conventional weighting gains G_conv in the high band. For the evaluation in this paper, the well-known Wiener filter [1] as well as the super-Gaussian joint MAP (JMAP) estimator [3] are used. In this investigation, the proposed noise suppression techniques with the use of α_opt and $\bar{\alpha}$ are thereby compared with the conventional case, where only the Wiener filter or only the JMAP

Parameter	Settings
Sampling frequency	16 kHz
Frame length L_F	160 ($\widehat{=}$20 ms due to downsampling)
FFT length M_F	256 (including zero-padding)
Frame overlap	50% (Hann window)
QMF filter length	64
Input SNR	-10 dB ... 35 dB (step size: 5 dB)
Noise estimation	Minimum Statistics [5]
SNR estimation	Decision-directed approach [2]
Number subbands M'_F	24
Number MFCCs N_C	13
Codebook size M_C	128 (training based on 1.5 h speech)

Table 2. System settings.

estimator is applied to both the low band and the high band. The parameters that have been used in the simulations are listed in Tab. 2. The look-up tables which are required for the estimation of α_opt were generated based on 10 min of clean speech from the NTT database disturbed by white Gaussian noise at different input SNR values.

In the simulation setup, the speech and noise signal can be filtered separately with weighting gains adapted for the noisy signal. Hence, the output signal can additionally be stated as $\hat{s}(k) = \tilde{s}(k) + \tilde{n}(k)$, where $\tilde{s}(k)$ is merely the filtered speech signal and $\tilde{n}(k)$ the filtered noise signal. Based on these quantities, the segmental speech SNR (SpSNR) and the segmental noise attenuation (NA) were calculated according to [10]. For the objective evaluation of the noise reduction schemes, seven speech signals from the NTT speech database were each degraded by four different noise types (cockpit, babble, factory1, buccaneer), taken from the NOISEX-92 database. Among the seven speech signals, there were four sequences from a male and three from a female speaker, each with a length of 8 seconds. The speech signals used for the evaluation were not included in the training data for the HMM and the look-up tables.

Figure 4 depicts the averaged results for SpSNR plotted over NA with the input SNR as control variable. Thus, a fair comparison with respect to the tradeoff noise attenuation and speech distortion is possible. The points of best performance would be placed in the upper right corner of the figure.

The objective measurements show that the additional use of the artificial BWE in the high band improves the results of conventional noise suppression techniques consistently. Especially at low input SNR values, where mainly the BWE approach is used (see Fig. 3), the new method outperforms either the Wiener filter or the JMAP estimator. As expected, the curves in Fig. 4 converge at higher input SNR values. Moreover, it can be seen that for the proposed method using $\bar{\alpha}$, the benefits of the noise attenuation are at the expense of a slightly lower segmental speech SNR for high SNR values. This can be explained by SNR estimation errors which lead to a suboptimal determination of the fading factor. Informal listening tests confirmed the instrumental measurements and showed that the occurrence of *musical tones* is reduced by the proposed method.

5. CONCLUSIONS

A novel approach to wideband speech enhancement has been presented in this paper that exploits spectral dependencies of speech signals. In an objective and subjective evaluation, it could be shown that the enhanced low band signal can be re-used to improve the results of a conventional noise suppression technique in the high band based on an artificial bandwidth extension. Although the computational complexity is increased by the new approach, the results motivate further investigations of this topic. In order to increase the perceived speech quality if only a noisy low band signal has been received, a slightly modified version of the system can additionally be used to perform a joint noise reduction and BWE.

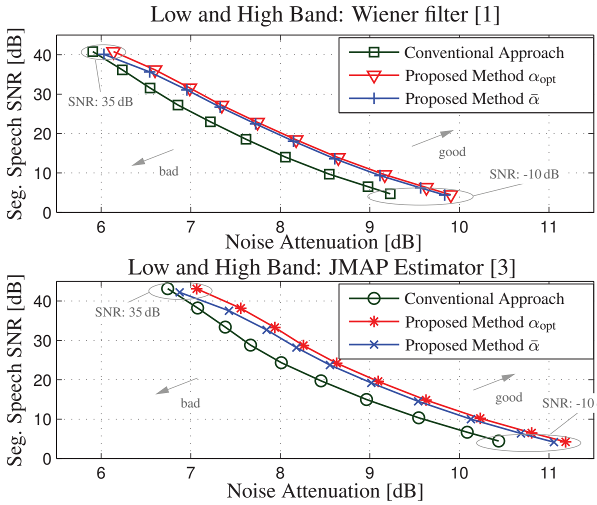

Fig. 4. Segmental speech SNR vs. noise attenuation.

6. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, pp. 1110–1126, 2005.

[4] C. Beaugeant, M. Schönle, and I. Varga, "Challenges of 16 kHz in Acoustic Pre- and Post-Processing for Terminals," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 98–104, May 2006.

[5] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 501–512, 2001.

[6] B. Geiser, H. Taddei, and P. Vary, "Artificial Bandwidth Extension without Side Information for ITU-T G.729.1," in *Proc. of INTERSPEECH*, Antwerp, Belgium, Aug. 2007.

[7] P. Jax and P. Vary, "Feature Selection for Improved Bandwidth Extension of Speech Signals," in *Proc. of ICASSP*, Montreal, Canada, May 2004.

[8] Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[9] P. Jax and P. Vary, "Bandwidth Extension of Speech Signals: A Catalyst for the Introduction of Wideband Speech Coding?" in *Communications Magazine, IEEE*, vol. 44, no. 5, May 2006, pp. 106–111.

[10] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A Psychoacoustic Approach to Combined Acoustic Echo Cancellation and Noise Reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 245–256, 2002.