SINGLE CHANNEL NOISE REDUCTION BASED ON AN AUDITORY FILTERBANK

Steffen Kortlang[1], Stephan D. Ewert[1], and Timo Gerkmann[2]

[1]Medical Physics Group, University of Oldenburg and Cluster of Excellence "Hearing4All", Oldenburg, Germany
[2]Speech Signal Processing Group, University of Oldenburg and Cluster of Excellence "Hearing4All", Oldenburg, Germany

ABSTRACT

Many noise reduction algorithms are designed in the short-time Fourier transform (STFT) domain. The STFT analysis results in frequency bands with a constant bandwidth. In contrast, perceptually motivated analysis-resynthesis filterbanks, such as the Gammatone filterbank, result in a higher frequency resolution in low frequencies as compared to high frequencies. This variable frequency resolution goes along with a changed temporal resolution and thus potentially different temporal correlations in the different frequency bands. In this paper, we design a noise power spectral density estimator at the output of a Gammatone filterbank. For this, we employ the state-of-the-art speech presence probability based estimator [13]. While [13] was designed in the STFT domain, in this paper the parameters of [13] are adjusted based on a statistical analysis of the changed temporal correlation at the output of the Gammatone filters. The proposed approach yields a comparable instrumentally predicted quality as the STFT-based baseline approach and thus allows for the integration of noise reduction with other algorithms that work in a perceptually motivated spectral domain.

Index Terms— speech enhancement, noise reduction, gammatone filter, auditory model

1. INTRODUCTION

Single-channel noise reduction algorithms (NRA) are often formulated in the frequency domain, e.g. using the short-time discrete Fourier transform (STFT) due to, among other reasons, the presence of computationally efficient implementations. Clean speech is estimated from the noisy speech by applying a time-variant gain function to the STFT coefficients. In the simplest approaches, a complex Gaussian distribution is assumed for the clean speech and noise STFT coefficients, resulting in the Wiener filter [19] as the minimum mean square error (MMSE)-optimal estimator of the complex clean speech coefficients or Ephraim and Malah's [6] MMSE-optimal estimator of the STFT-amplitudes. Spectral noise power estimators are often based on minimum statistics [22]. In contrast, [12, 13] propose a noise power estimator based on the speech-presence-probability (SPP) in each time-frequency point. Estimation of the spectral speech power can be performed by, e.g., employing the decision-directed approach [6] or cepstral smoothing techniques [3].

It was shown that adequate spectral smoothing is a necessary step for high-quality noise reduction systems, e.g. [2, 3, 7]. This smoothing can be obtained, for instance, by averaging adjacent spectral coefficients. In contrast to conventional STFT-based approaches, the transformation of the input signal to the spectral domain is performed by applying a filterbank with fewer frequency channels. Several analysis-synthesis filterbanks have been used successfully in speech enhancement applications, like the low-delay filterbanks with adaptive subband filtering proposed by [21], which can achieve a comparable quality of enhanced speech but with lower signal delay. For our proposed noise reduction scheme, we use an auditorily-motivated analysis-resynthesis filterbank based on Gammatone filters [23] with a low-delay and nearly perfect signal reconstruction [17]. Such a system closely mimics the auditory spectro-temporal resolution and potentially offers perceptual advantages regarding the time-frequency distribution of artifacts originating from NRA. The noise power estimation is implemented analogous to the algorithm proposed in [12, 13]. For this, the STFT-based a posteriori SPP estimation and its proposed parameter optimization [11] are adapted to the GFB analysis system. The proposed multiband noise reduction scheme can be easily combined with, e.g., state-of-the-art dynamic compressors using an auditory filterbank [8, 9].

2. ALGORITHM OVERVIEW

2.1. Analysis, resynthesis and frame processing

As applied in an auditory perception model [5], 4th-order linear Gammatone filters can be used to mimic human auditory filters. A Matlab implementation of the Gammatone filterbank (GFB) with the analysis-synthesis system proposed in [17] was used in the proposed algorithm. The GFB was realized using cascaded first-order complex-valued bandpass filters in an all-pole design, resulting in computationally efficient infinite impulse response bandpass filters.

978-1-4799-6808-4/14/$31.00 ©2014 IEEE

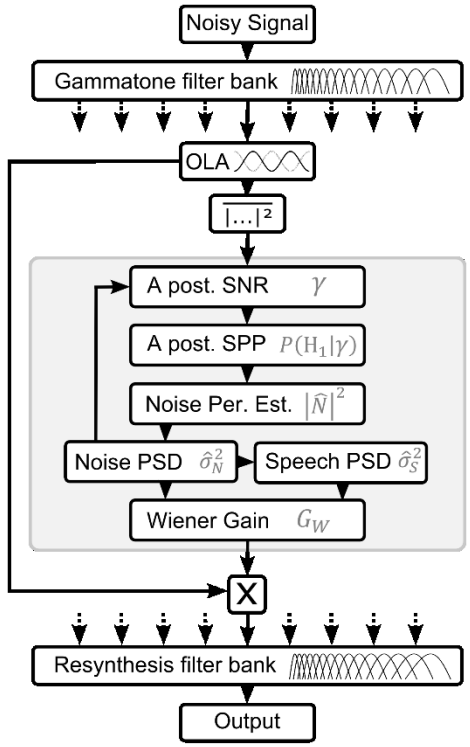

Fig. 1. Block diagram of the proposed noise reduction algorithm with Gammatone filterbank analysis and resynthesis.

The block diagram of the algorithm is shown in Fig. 1. The input signal (at a sampling rate of 16 kHz) is processed by a 4th-order Gammatone analysis filterbank. The center frequencies are linearly spaced on the equivalent rectangular bandwidth (ERB) frequency scale [14]. The filter bandwidths were set to 1 ERB resulting in 33 bands with center frequencies ranging from 26.1 to 7792.7 Hz and bandwidths from 27.5 to 864.5 Hz, approximately overlapping at their 3-dB point. In each GFB channel k, the absolute value of the complex-valued output represents an approximation of the bandpass filtered Hilbert envelope. After GFB analysis, Hann-windows $w(n)$ with a constant duration of T_w = 32 ms (N = 512 samples), where successive segments overlap by 50%, were used in an overlap-add (OLA) manner. In this way, all estimates in the subsequent stages are updated each 16 ms. This approach is referred to as GFB1 in the following. In an alternative approach, GFB2, the frame duration was inversely related to the filter bandwidth (see Sec. 3.1. for details).

The resulting Wiener Gain G_W from the NRA system (gray block in Fig. 1) is applied to each segment. After the segments are added up again in each channel, a 2nd-order Gammatone resynthesis filterbank is applied to attenuate frequency components outside the desired channel and thus reduce distortions. Before summation and taking the real value as output, all channels were time aligned to compensate for their different group delays, resulting in an overall delay of 6.25 ms (corresponding to 100 samples; see [16, 17] for details).

2.2. Noise and speech power estimation

The noise and speech power estimation stage is indicated by a gray background in Fig. 1. The noise power estimation is realized analogously to [12, 13], but adapted to the GFB output. In each GFB channel k and windowed segment l, a short-term periodogram-like energy estimate of the (noisy) input signal Y is calculated as

$$|Y_k(l)|^2 = \frac{1}{N}\sum_n \big(|g_k(n)| \cdot w(n)\big)^2, \quad (1)$$

where n denotes the segment sample index, N denotes the total segment length and g_k denotes the complex GFB output of channel k, whose absolute value (Hilbert envelope) is multiplied with the Hann-window $w(n)$. It is assumed that the speech (S) and the noise (N) power are additive: $|Y_k(l)|^2 = |S_k(l)|^2 + |N(l)|^2$. The bandpass filtered noise and speech power is defined as the expected value of the corresponding periodogram estimate $\hat{\sigma}_S^2 = E\{|S|^2\}$ and $\hat{\sigma}_N^2 = E\{|N|^2\}$, respectively.

Introducing the a posteriori signal-to-noise ratio (SNR), $\gamma = |Y|^2/\hat{\sigma}_N^2$, the probability of speech presence (SPP) given the observation γ, can be written as a function of the generalized likelihood ratio (GLR), Λ:

$$P(\mathrm{H}_1|\gamma) = \frac{\Lambda}{1+\Lambda}, \quad (2)$$

where Λ is defined as the ratio of the likelihood of speech presence and the likelihood of speech absence, weighted by their prior probabilities of speech presence:

$$\Lambda(\gamma) = \frac{P(\mathrm{H}_1)\, p(\gamma|\mathrm{H}_1)}{P(\mathrm{H}_0)\, p(\gamma|\mathrm{H}_0)}, \quad (3)$$

where $P(\mathrm{H}_1)$ is the a priori SPP, $P(\mathrm{H}_0) = 1 - P(\mathrm{H}_1)$, and $p(\gamma|\mathrm{H}_1)$ and $p(\gamma|\mathrm{H}_0)$ are the likelihoods of speech presence and absence. Assuming that the real and imaginary parts of the complex GFB output are Gaussian distributed, the likelihoods of the a posteriori SNR follow a χ^2-distribution. With these assumptions, equation (3) results in:

$$\Lambda(\gamma) = \frac{P(\mathrm{H}_1)}{P(\mathrm{H}_0)} \cdot \left(\frac{1}{1+\xi_{opt}}\right)^{r/2} \exp\left(\frac{\xi_{opt}}{1+\xi_{opt}} \frac{r}{2} \gamma\right), \quad (4)$$

where the parameter ξ_{opt} reflects the SNR that can be expected in speech presence, and the parameter r is referred to as the degrees of freedom of the χ^2-distribution.

By averaging the Gammatone filter outputs in frames of 32 ms (see Fig. 1), the degrees of freedom r increase. The amount of increase depends on the correlation of the data that is averaged [11]. The larger the correlation, the higher the increase in the degrees of freedom r. The parameter ξ_{opt} and r were fitted and optimized to the proposed GFB analysis system as described in [11] in section 2.3.

For each frame in each channel, the a posteriori speech presence probability (SPP) is calculated according to Eqs. (2, 4) using the noise power estimate of the previous frame for the calculation of a posteriori SNR. Thereafter, the noise periodogram estimate $\left|\widehat{N}\right|$ is updated as

$$\left|\widehat{N}\right|^2 = P(\mathrm{H}_0|Y)\,|Y|^2 + P(\mathrm{H}_1|Y)\,\hat{\sigma}_N^2. \quad (5)$$

The spectral noise power estimate $\hat{\sigma}_N^2$ is then obtained by temporal smoothing:

$$\hat{\sigma}_N^2(n) = \alpha_N \cdot \hat{\sigma}_N^2(n-1) + (1-\alpha_N)\left|\widehat{N}(n)\right|. \quad (6)$$

In each channel, the recursive smoothing factors α_N are set to correspond to a time constant of 64 ms (here: 0.8 for a parameter update rate of 16 ms). The final speech power estimation for noise reduction purpose consists of three parts. First, the noise power spectral density (PSD) is estimated as previously described. Secondly, a decision-directed approach [6] is used to get an estimate of the speech PSD:

$$\hat{\sigma}_S^2(n) = \alpha_S\,|\hat{S}(n-1)|^2 + (1-\alpha_S)\,\max\{|Y|^2 - \sigma_N^2, 0\}. \quad (7)$$

As a first approximation, the filter coefficients α_S are set to correspond to an equivalent time constant of 784 ms of a 1st-order lowpass filter (here: 0.98 for a parameter update rate of 16 ms). Thirdly, a Wiener gain G_W is calculated, limited to G_{min} = -12 dB and applied as follows:

$$\left|\hat{S}\right|^2 = {G_W}^2|Y|^2 = \left(\frac{\hat{\sigma}_S^2}{\hat{\sigma}_S^2+\hat{\sigma}_N^2}\right)^2 |Y|^2. \quad (8)$$

2.3. Optimal parameters for SPP estimation

In each channel, the degrees of freedom r were estimated according to [11]. For this, Gaussian white noise is created with zero mean and variance (VAR) of 1. With the a posteriori SNR, defined as

$$\gamma_k(n) = |Y_k(n)|^2/\hat{\sigma}_{N,k}^2, \quad (9)$$

the degrees of freedom are obtained as [11]

$$\bar{r}_k = \frac{2}{\mathrm{VAR}\{\gamma_k\}}. \quad (10)$$

As a consequence of the frequency-dependent bandwidth of the GFB channels, the resulting degrees of freedom are channel dependent (proportional to the filter bandwidth of 1 ERB). The resulting values are given in the middle column of Tab. 1 (GFB1) ranging from from 2.5 to 55.9.

The assumed a priori SNR ξ_{opt} is chosen to obtain a specified performance in terms of false alarms and missed detections for a given range of input SNRs between −10 and 15 dB, as detailed in [11]. The resulting ξ_{opt} ranges from 11 dB for the lowest channel up to 3 dB for the highest channel.

3. EVALUATION

3.1. Algorithms and stimuli

For further analysis, three different approaches were considered: GFB1 had a fixed frame duration of 32 ms. In GFB2, the frame durations were selected in such a way that the a posteriori SNRs at the output of the Gammatone filters exhibit the same variance and thus the same degrees of freedom. For this purpose, the frame durations were set to 5 times the inverse filter bandwidth in Hz (from 182 ms to 6 ms; see Table 1). In addition to the GFB approaches, a conventional STFT-based algorithm was tested. A square-root Hann-window of length T_w = 32 ms with 50% overlap was applied to each frame prior to a DFT analysis and after application of the inverse DFT. The a posteriori signal-to-noise ratio is smoothed as proposed by [11] to reduce the variance of the SPP estimate and thus decrease local distortions. The smoothing spans a range of 105.5 Hz along frequency and 64 ms along time. For the STFT and GFB2 approach, the values of r were averaged over the frequency channels, and resulted in $r = 11$, respectively (see Table 1). For ξ_{opt}, a value of 8 dB for both STFT and GFB2 was obtained. The a priori SPP was set to $P(\mathrm{H}_1) = 0.5$.

In further evaluations, results are given as the average of 20 sentences from the TIMIT database [10] (10 male, 10 female), for input SNRs between -10 dB and 15 dB. Noise signals were stationary white Gaussian noise and babble noise in a cafeteria taken from the NOISEX-92 database [24].

	STFT	GFB1	GFB2
T_w	32 ms	32 ms	(182…6) ms
r	11	2.5 … 55.9	11
ξ_{opt}	8 dB	(11 … 3) dB	8 dB

Tab. 1. Parameter settings for the three considered algorithms.

3.2. SNR improvement

In addition to the noisy signal, the same speech enhancement Wiener gains are applied block-wise to the speech-only and the noise-only signals. This linear filtering procedure is referred to as shadow filtering. The segmental noise reduction (NR), as well as the segmental speech SNR (spSSNR) as proposed by [20] were used for evaluation. The resynthesized time-domain signals were segmented into non-overlapping 10-ms segments. To focus on the noise reduction during speech, only signal frames with energy larger than −45 dB compared to the maximum frame energy were considered. While NR is a measure of the relative noise reduction, spSSNR takes into account undesired speech distortions and becomes larger the lower the speech distortions are.

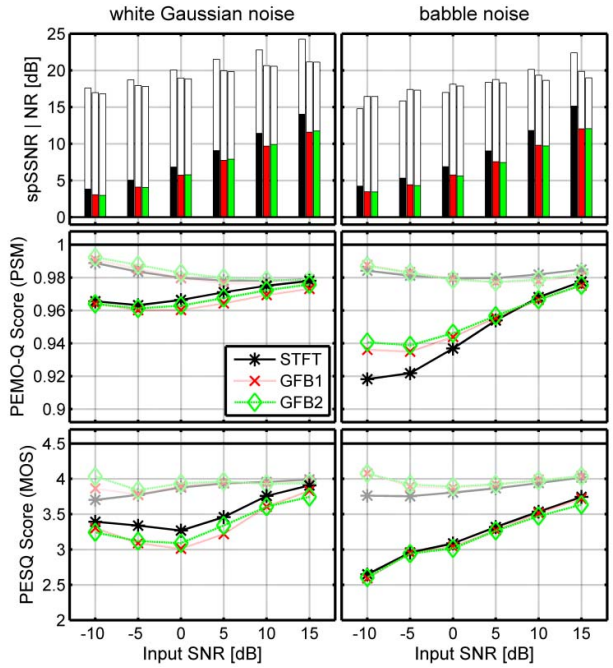

Fig. 2. Simulation results as a function of the input SNR for white Gaussian noise (left) and babble noise (right). Top: Segmental speech SNR (filled bars) and the segmental noise reduction (white bars). The mean standard deviations (msd) of the 20 TIMIT sentences averaged across all input SNRs amount to 1.1 dB and 1.5 dB, respectively. Middle: PEMO-Q Score (PSM; msd: 0.0045). Bottom: PESQ Score (MOS; msd: 0.20). The light lines in the lower two panels refer to the speech-based quality measures (spPSM and spMOS with a msd of 0.0019 and 0.091, respectively).

The upper panels of Fig. 2 show the results for stationary white Gaussian noise (left) and babble noise (right) as a function of the input SNR for spSSNR (filled bars) and NR (white bars). The total height of both bars is thus a measure of the overall performance. For both noises, the STFT algorithm reaches higher values of the spSSNR because of the higher frequency resolution. Regarding the NR, the 3 NRAs perform similar in white noise. However, in babble noise, the GFB approaches reach higher noise reductions at low SNRs compared to the STFT implementation. Only low to negligible influence of the adapted window length (GFB2) can be observed.

3.3. Instrumental prediction of perceived quality

The perceived audio quality was predicted based on two reference-based instrumental audio quality measures. PEMO-Q [18] evaluates the similarity between internal representations of the noisy and reference audio signal after applying an auditory perception model [5]. The resulting correlation is referred to as Perceptual Similarity Measure (PSM). The second measure was the ITU standardized (ITU-T P.862) PESQ ("Perceptual Evaluation of Speech Quality", [1]). The PESQ score is mapped to a mean opinion score (MOS), which gives values between 0.5 and 4.5. Both measures have been shown to have reasonably high correlations with the perceived quality of noise-reduced speech [15]. PSM and MOS values were assessed using the processed noisy speech signal as a target signal and a noisy input signal with a 12-dB improved SNR (equal to G_{min}) as reference. Additionally, the measures were applied to the shadow-filtered speech, where the same Wiener gains were applied block-wise to the clean speech alone. Here, the clean speech signal was used as reference. The respective outcome is referred to as spPSM and spMOS, respectively, and may predict perceptual speech distortions.

The quality predictions are given in Fig. 2 for the white Gaussian noise (left panels) and the babble noise (right panels) as a function of the input SNR. Results from the PEMO-Q scores are given in the medium panels (PSM in dark and spPSM in light colors) while the PESQ scores are given in the lower panels (MOS in dark and spMOS in light colors). While in stationary noise the STFT approach appears to slightly outperform the GFB algorithms in terms of audio quality, the PEMO-Q score indicates improved audio quality for the GFB algorithms in babble at low SNRs. Speech distortions are generally similar or even better (at low SNRs).

4. CONCLUSION

In this work, we present an algorithm for single channel speech enhancement at the output of the auditorily-motivated Gammatone filterbank. The employed Wiener filter based single channel speech enhancement algorithm requires an estimate of the noise power spectral density. This noise power spectral density can for instance be estimated based on the a posteriori speech presence probability. In this paper, we optimize the statistical parameters of the speech presence probability estimator to the different temporal correlation at the output of individual Gammatone channels. The optimization is done by interpreting the speech presence probability estimator as a detector and minimizing the missed hit and false alarm rates in each channel. The resulting speech presence probability estimator was then used to estimate the noise power spectral density employed in Wiener filtering.

Two Gammatone-based approaches were compared to an STFT approach. The proposed systems show lower segmental speech SNR values. However, an increase in the segmental noise reduction can be achieved in babble noise. All algorithms perform comparably in instrumentally predicted sound quality with a small benefit for the Gammatone based approaches in strong babble noise. The proposed noise reduction scheme provides the capability for an integration into other perceptually motivated multiband systems.

This work was supported by BMBF 13EZ1127D and DFG FOR 1732 (TPE).

5. REFERENCES

[1] J. G. Beerends, A. P. Hekstra, A. W. Rix and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ): The New ITU standard for end-to-end speech quality assessment Part II. Psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp.765–778, 2002.

[2] M. Brandt and J. Bitzer, "Optimal Spectral Smoothing in Short-Time Spectral Attenuation (STSA) Algorithms: Results of Objective Measures and Listening Tests," in *17th Eur. Signal Process. Conf (EUSIPCO)*. 2009, pp. 199–203.

[3] C. Breithaupt, T. Gerkmann and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2008, pp. 4897–4900.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[5] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation: I. detection and masking with narrow-band carriers," *J. Acoust. Soc. Amer.*, vol. 102, no. 5, pp. 2892–2905, Nov. 1997.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[7] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement system," in *Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2009, pp. 4409–4412.

[8] S. D. Ewert and G. Grimm, "Model-based hearing aid gain prescription rule," in *Speech perception and auditory disorders, Int. Symposium on Audiological and Auditory Research (ISAAR)*, Aug. 2011.

[9] S. D. Ewert, S. Kortlang and V. Hohmann, "A model-based hearing aid: Psychoacoustics, models and algorithms," in *Proceedings of Meetings on Acoustics (POMA)*, vol. 19, 2013.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett and N. L. Dahlgren, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM," National Institute of Standards and Technology (NIST), 1990.

[11] T. Gerkmann, C. Breithaupt and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Language Process.,* vol. 16, no. 5, pp. 910–919, Jul. 2008.

[12] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Workshop Appl. Signal Process. Audio, Acoust*. IEEE, 2011, pp.145–148.

[13] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based Noise Power Estimation with Low Complexity and Low Tracking Delay," *IEEE Trans. Audio, Speech, Language Process.,* vol. 20, no. 4, pp. 1383–1393, May *2012*.

[14] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res*., vol. 47, pp. 103–138, Feb. 1990.

[15] N. Harlander, R. Huber and S. D. Ewert, "Sound Quality Assessment Using Auditory Models," *J. Audio Eng. Soc.*, vol. 62, no. 5, pp. 324–336, 2014.

[16] T. Herzke and V. Hohmann, "Improved numerical methods for gammatone filterbank analysis and synthesis," *Acta. Acust. Acust.*, vol. 93, pp. 498–500, 2007.

[17] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta. Acust. Acust.*, vol. 88, pp. 433–442, 2002.

[18] R. Huber and B. Kollmeier, "PEMO-Q – A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Trans. Audio, Speech, Language Process.,* vol. 14, no. 6, pp. 1902–1911, 2006.

[19] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[20] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Applied Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, Jan. 2005.

[21] H. W. Löllmann and P. Vary, "Uniform and warped low delay filer-banks for speech enhancement," *J. Speech Comm.*, vol. 49, no. 7-8, pp. 574–587, 2007.

[22] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[23] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth and P. Rice, "An efficient auditory filterbank based on the gammatone function," Paper presented at a meeting of the IOC Speech Group on Auditory Modelling at RSRE, Dec. 1987.

[24] A. P. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *J. Speech Comm.*, vol. 12, no. 3, pp. 247–251, 1993.