A Model-Based Spectral Envelope Wiener Filter for Perceptually Motivated Speech Enhancement

Najib Hadir, Friedrich Faubel, Dietrich Klakow

Spoken Language Systems,
Saarland University, D-66123 Saarbrücken, Germany
{najib.hadir,friedrich.faubel,dietrich.klakow}@lsv.uni-saarland

Abstract

In this work, we present a model-based Wiener filter whose frequency response is optimized in the dimensionally reduced log-Mel domain. That is achieved by making use of a reasonably novel speech feature enhancement approach that has originally been developed in the area of speech recognition. Its combination with Wiener filtering is motivated by the fact that signal reconstruction from log-Mel features sounds very unnatural. Hence, we correct only the spectral envelope and preserve the fine spectral structure of the noisy signal. Experiments on a Wall Street Journal corpus showed a relative improvement of up to 24% relative in PESQ and 45% relative in log spectral distance (LSD), compared to Ephraim and Mallah's log spectral amplitude estimator.

Index Terms: speech enhancement, Bayesian estimation, signal reconstruction

1. Introduction

It is well-known that the presence of background noise can significantly degrade the quality and intelligibility of speech. Next to direct consequences, such as listener fatigue, it can cause further degradations in subsequent processing steps. This in particular concerns LPC based speech coders, which are optimized for operation in a noise-free environment. Hence, the development of such devices in the early 1970s also triggered serious research efforts into speech enhancement methods, starting with Weiss's spectral subtraction [1], Calahan's Wiener filter [2] as well as Lim and Oppenheim's iterative variant thereof [3]. The aim of all these methods is to improve the subjective quality of speech, and, nonetheless, they also introduce filtering artifacts such as "musical noise". This can be attributed to the facts that (1) the optimization is not done in a domain that is perceptually relevant; that (2) it is extremely difficult to get accurate estimates of the short-time (power) spectral densities of clean speech and noise; and that (3) the effect of the relative phase between speech and noise is disregarded.

More recent approaches have tried to address these points by working in the perceptually more relevant logarithmic spectral magnitude domain [4, 5, 6], by exploiting the super-Gaussian character of speech [7] or by using strong prior models for how clean speech looks like [8, 9, 10, 5, 6]. Especially Burshtein and Gannot [5] as well as Nilsson et al. [6] took an interesting new approach, which employs noise compensation methods that were originally developed in the area of automatic speech recognition. In this paper, we further pursue this approach by using a spectral domain Wiener filter whose frequency response is determined based on a Gaussian mixture model (GMM) of the clean speech distribution. The novel aspects of the work are in particular that

1. we optimize the frequency response of the filter in the logarithmic Mel (log-Mel) domain and thereby correct the perceptually relevant part of the spectral envelope rather than the spectral fine structure.
2. we determine the frequency response of the Wiener filter based on a very recent minimum mean square error (MMSE) speech feature enhancement approach [11, 12] instead of using the log-normal approximation [6].
3. we model the effect of noise to log-Mel speech spectra under consideration of the relative phase [13].

The performed experiments show significant gains in non-stationary noise environments over state-of-the-art implementations [14] of spectral subtraction, Wiener filtering and Ephraim and Mallah's log-spectral amplitude estimator. The remaining part of this paper is organized as follows. Section 2 briefly reviews the Bayesian speech feature enhancement approach taken in [11, 12]. All the variables in that section denote log-Mel speech spectra. The operations are performed component-wise. Section 3 explains how the clean speech signal can be reconstructed either by direct re-synthesis or by application of a Wiener filter. Section 4 finally presents experimental results on the MC-WSJ-AV corpus [15], with different types of noise from the NOISEX-92 database [16] added at various signal-to-noise ratios.

2. Bayesian Speech Feature Enhancement

The speech feature enhancement approach taken in [11, 12] aims at suppressing noise in log-Mel speech features based on a Gaussian mixture model of the clean speech distribution:

$$p(\mathbf{x}) = \sum_{k=1}^{K} c_k \mathcal{N}(\mathbf{x}; \mu_{X|k}, \mathbf{\Sigma}_{X|k}). \quad (1)$$

In this equation, c_k, $\mu_{X|k}$ and $\mathbf{\Sigma}_{X|k}$ denote the prior probability, mean and covariance matrix, respectively, of the k-th Gaussian component. Then, further modeling background noise $\mathbf{N}$ as a Gaussian random variable with distribution $p(\mathbf{n}) = \mathcal{N}(\mathbf{n}; \mu_N, \Sigma_N)$, the joint distribution of clean speech X and noisy speech Y can be obtained through the following transformation of random variables:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \tilde{f}\left(\begin{bmatrix} X \\ N \end{bmatrix}\right) \quad \text{with} \quad \tilde{f}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{n} \end{bmatrix}\right) = \begin{bmatrix} \mathbf{x} \\ f(\mathbf{x}, \mathbf{n}) \end{bmatrix} \quad (2)$$

This work has been supported by the Federal Republic of Germany, through the Cluster of Excellence for Multimodal Computing and Interaction (MMCI).

where $f(\mathbf{x},\mathbf{n}) = \mathbf{x}+\mathcal{E}_{p(\alpha)}\{\log(1+e^{\mathbf{n}-\mathbf{x}}+2\alpha\sqrt{e^{\mathbf{n}-\mathbf{x}}})\}$ is the phase-averaged interaction function from [13], which describes the MMSE relationship between noisy speech, clean speech and noise in the log-Mel domain. In this work, the transformation in (2) is approximated with a sequence of unscented transforms, as described in [11, 17]. The result is a joint Gaussian mixture,

$$p\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right) = \sum_{k=1}^{K} c_k \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}; \begin{bmatrix}\mu_{X|k}\\ \mu_{Y|k}\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma}_{X|k} & \mathbf{\Sigma}_{XY|k}\\ \mathbf{\Sigma}_{YX|k} & \mathbf{\Sigma}_{Y|k}\end{bmatrix}\right),$$

with which the MMSE estimate $\widehat{\mathbf{x}}_i$ of clean speech given noisy speech $\mathbf{y}_i$ can be approximated according to [18, 12]:

$$\begin{aligned}\widehat{\mathbf{x}}_i &= \int \mathbf{x}_i p(\mathbf{x}_i|\mathbf{y}_i) d\mathbf{x}_i \\ &\approx \sum_{k=1}^{K} p(k|\mathbf{y}_i)\big(\mathbf{y}_i - \underbrace{(\mu_{Y|k} - \mu_{X|k})}_{\Delta_k}\big). \end{aligned} \quad (3)$$

We call this approximation mode-dependent bias correction (MDBC) as it corrects (subtracts) the bias Δ_k introduced to each mode weighted with the posterior probability $p(k|\mathbf{y}_i)$ of being in that mode. Following [18, 12], $p(k|\mathbf{y}_i)$ is evaluated with Bayes rule:

$$p(k|\mathbf{y}_i) = \frac{c_k \mathcal{N}(\mathbf{y}_i; \mu_{Y|k}, \mathbf{\Sigma}_{Y|k})}{\sum_{k'=1}^{K} c_{k'} \mathcal{N}(\mathbf{y}_i; \mu_{Y|k'}, \mathbf{\Sigma}_{Y|k'})}. \quad (4)$$

The required noise distribution $p(\mathbf{n})$ can be estimated with the expectation maximization algorithm, as explained in [17, 12]. Alternatively, the joint distribution can be learned from a joint clean / noisy speech training corpus [18].

3. Clean Speech Reconstruction

The Bayesian speech feature enhancement approach from the previous section provides a minimum mean square error (MMSE) estimator for log-Mel clean speech features. The resulting, cleaned features are typically passed on to the speech recognizer, possibly after application of a discrete cosine transform as well as further processing steps such as cepstral mean and variance normalization [18, 11, 17, 12]. In this work, however, we are interested in re-synthesizing the speech signal. Hence, we follow the approaches taken in [5] and [6] and either invert all the feature extraction steps up to the magnitude domain and then multiply by the noisy phase; or use the estimated clean speech features for constructing the frequency response of an adaptive Wiener filter. Both these approaches are explained in more detail in the following.

3.1. Direct Reconstruction from Log-Mel Features

Similar as proposed in [5] for log magnitude spectra, the speech signal can be reconstructed from log-Mel features by inverting the feature extraction steps and subsequently multiplying by the original phase of the noisy speech signal. For log-Mel features, the feature chain consists of a short-time Fourier transform (FFT), followed by a component-wise magnitude square, multiplication by an $M \times N$ triangular Mel filterbank matrix $\mathbf{W}$ and, finally, logarithmic compression. Put into equations, log-Mel features are calculated according to:

$$\mathbf{x}_i = \log\left(\mathbf{W} \cdot P_{X_i}\right) \quad \text{with} \quad P_{X_i} = \left\|\text{FFT}\left\{\mathbf{x}_i^{sig}\right\}\right\|^2$$

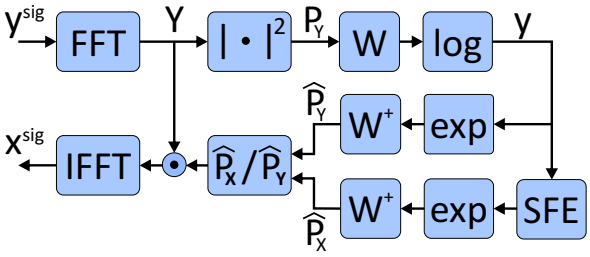

Figure 1: *Block diagram of the model-based Wiener filter with speech feature enhancement (SFE).*

where $\mathbf{x}_i^{sig}$ denotes the i-th block of samples and where P_{X_i} denotes the power spectral density. Conversely, given an estimated clean speech feature $\widehat{\mathbf{x}}_i$ in the log-Mel domain, the corresponding power spectral density $\widehat{P}_{X_i}$ can be approximated as:

$$\widehat{P}_{X_i} = \mathbf{W}^{\dagger} \cdot \exp(\widehat{\mathbf{x}}_i) \quad (5)$$

where $\mathbf{W}^{\dagger}$ denotes the $N \times M$ Moore-Penrose pseudoinverse of $\mathbf{W}$. Hence, the sample block $\mathbf{x}_i^{sig}$ can be reconstructed as

$$\mathbf{x}_i^{sig} = \text{IFFT}\left\{\sqrt{\widehat{P}_{X_i}} \cdot e^{j\arg(Y_i)}\right\},$$

where IFFT denotes the inverse Fourier transform and where $\arg(Y_i)$ is the phase spectrum of the observed noisy speech signal. The reconstructed signal is obtained by overlapping and adding the $\mathbf{x}_i^{sig}$.

The use of log-Mel features has several advantages. Firstly, log-Mel features emulate both the logarithmic frequency and intensity perception of the human auditory system and thereby operate in a domain that is perceptually relevant. Secondly, they capture the spectral envelope and thereby the formants, which are most important for intelligibility. Thirdly, their distribution is easier to learn, as they are dimensionally reduced and as they are more Gaussian (due to the averaging). All these advantages are, however, matched by the fact that they also destroy the fine spectral structure, especially if $M << N$. This introduces severe degradations of the sound quality.

3.2. Model-Based Wiener Filtering

In order to avoid the unnatural, synthetic sounding voice, which results from direct reconstruction, we decided to use the clean speech estimate $\widehat{\mathbf{x}}_i$ from Section 2 for driving a Wiener filter. That is achieved by taking the clean speech power spectral density estimate $\widehat{P}_{X_i}$ from (5) and then calculating the frequency response of the Wiener filter according to:

$$H_i(\omega) = \frac{\widehat{P}_{X_i}(\omega)}{\widehat{P}_{X_i}(\omega) + \widehat{P}_{N_i}(\omega)} = \frac{\widehat{P}_{X_i}(\omega)}{\widehat{P}_{Y_i}(\omega)}, \quad (6)$$

as shown in the diagram in Figure 1. Note, in particular, that we approximate the power spectral density $\widehat{P}_{Y_i}$ of the noisy signal as $\mathbf{W}^{\dagger} \cdot \exp(\mathbf{y}_i)$. This is an important detail, as the use of this smoothed version of the original power spectral density P_{Y_i} yields a spectral envelope Wiener filter that preserves the fine spectral structure. The resulting frequency response is shown in Figure 2, with the surface plot exemplifying the filters variation in time over 60 frames of a noisy utterance.

It is noteworthy to mention that Nilsson et al. [6] took a similar approach. In that work, however, the power spectral

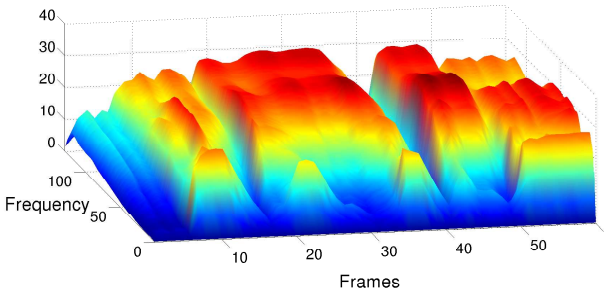

Figure 2: *Example of the frequency response of the proposed Wiener filter, plotted in* $\log(H(\omega)+1)$ *for better visibility.*

shift	16	32	64	128	256
PESQ	2.474	2.4860	2.432	2.358	1.886

Table 1: *Effect of the shift factor on the PESQ, for model-based Wiener filtering of the MC-WSJ-AV corpus [15] contaminated with 5dB factory(2) noise from the NOISEX-92 database [16].*

densities P_{X_i} and P_{N_i} of clean speech and noise were obtained with parallel model combination, under a log-normal assumption. Spectral envelope filtering was not performed as the approach operates in the log-spectral domain. The effect of the relative phase was ignored.

3.3. Implementation Details

In our implementation of model-based Wiener filtering, all the processing was done at 16kHz. For calculating the short-time Fourier transform, we cut the signal into frames of 256 samples and then applied a Hamming window. These frames were then further processes as described in Section 3.2, under consideration of the following implementation details.

Frame Shift: Instead of using a frame-shift of 160 samples[1], as in our speech recognition system, we investigated how different frame-shifts affect the sound quality. Informal listening tests seemed to indicate that the smaller the frame shift the less was the amount of crackling present in the processed audio file. This was confirmed by computing PESQ scores [14] (shown in Table 1). As the sound quality did not further improve with frame-shifts smaller than 32 samples, we used this number in all of the following experiments.

Mel Filterbank: As a Mel filterbank we used 30 triangular shaped Mel filters, which covered a total frequency range of 0-7kHz. The pseudo-inverse of the Mel filterbank matrix was calculated via singular value decomposition.

Smoothing: In order to avoid strong inter-frame fluctuations of the frequency response H_i, we further smoothed the power spectral densities P_{X_i} and P_{Y_i} in time. This was achieved by first-order recursive averaging,

$$\begin{aligned} \overline{P}_{X_i} &= \alpha \overline{P}_{X_{i-1}} + (1-\alpha) P_{X_i}, \\ \overline{P}_{Y_i} &= \alpha \overline{P}_{Y_{i-1}} + (1-\alpha) Y_{X_i}, \end{aligned}$$

with smoothing factor α. Based on the experimental results from Figure 3, we chose a smoothing factor of $\alpha = 0.85$.

[1] this corresponds to a window overlap of 62.5%

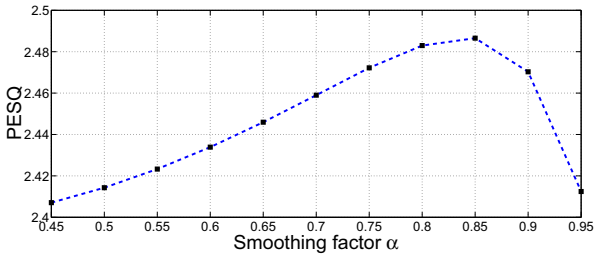

Figure 3: *Effect of the smoothing factor* α *on the PESQ, under the same conditions as in Table 1.*

4. Experiments

In order to evaluate the performance of the proposed Wiener filter under controlled conditions, we performed a series of experiments on the close-talking channel of the Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus [15]. This corpus consists of 352 utterances spoken by 10 speakers, with a total recording length of 40 minutes. Different noise conditions were simulated by adding noise from the NOISEX-92 database [16] at different signal-to-noise ratios (SNRs). The performance of the proposed speech enhancement algorithms was evaluated by calculating perceptual evaluation of speech quality (PESQ) scores [14] as well as the log spectral distance (LSD) [14] between clean and estimated clean speech spectra.

Figure 4 shows the PESQ scores we obtained for noisy and enhanced speech. The noise conditions range from relatively stationary car driving noise (volvo) over factory noise (factory2) up to highly non-stationary babble noise (babble), added with SNRs of 0, 5 and 10dB, respectively. Speech enhancement results are given for Berouti's spectral subtraction (SS) [19], Scalart and Filho's Wiener filter (WF) [20], [14, §6.10], Ephraim and Malah's MMSE log spectral amplitude estimator (log-MMSE) [4], the proposed model-based Wiener filter (SFE-WF) from Section 3.2 as well as the the direct reconstruction approach (SFE-REC) from Section 3.1. The bar plots clearly show that the proposed SFE-WF algorithm outperforms all the other methods in non-stationary noise. The largest gain was achieved for 0dB babble noise, with an improvement of 0.35 (24.0% relative) over log-MMSE and an improvement of 0.58 (47.2% relative) over SS. In stationary noise (volvo), the SFE-WF performed comparably to log-MMSE, with an average PESQ of 3.49 in both cases.

Figure 5 shows log-spectral distances (LSDs) for the same noise conditions and enhancement techniques that were used during calculation of the PESQ scores in Figure 4. When comparing these plots, however, it should be taken into account that improvements in PESQ mean higher scores wile improvements in LSD mean lower values. Hence, Figure 5 shows that the proposed model-based Wiener filter (SFE-WF) consistently outperformed the log-spectral amplitude estimator (log-MMSE) in all noise conditions. Another interesting result is that spectral subtraction performed surprisingly well. For babble noise at 0dB it even outperformed the SFE-WF, by 5.4% relative. With Volvo noise, however, spectral subtraction performed 50.0%, 39.0% and 29.6% relative worse than the SFE-WF.

Informal listening tests revealed that spectral subtraction and Wiener filtering suffered from musical noise [14], especially at lower SNRs. The model-based Wiener filter as well as the log spectral amplitude estimator, on the other hand, did not seem to be affected by this problem. But they introduced other

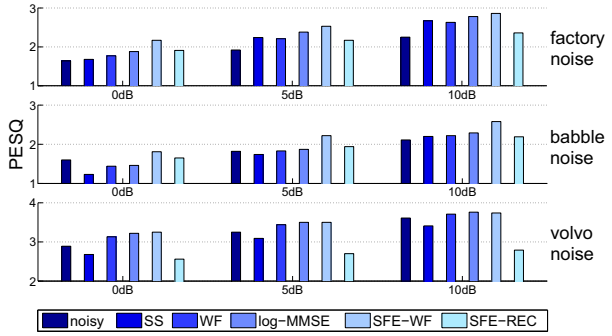

Figure 4: *PESQ scores for different noise types and enhancement techniques, averaged over the MC-WSJ-AV corpus. The best achievable score is 4.5. The worst is 1.0.*

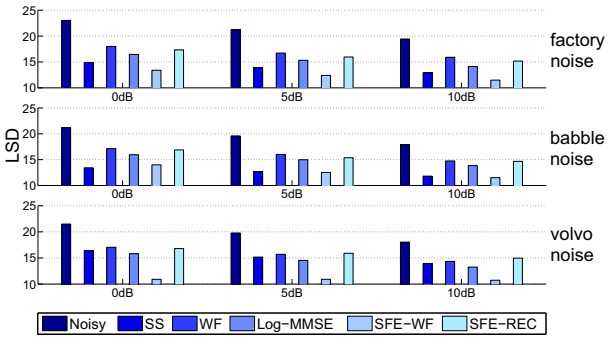

Figure 5: *Log-spectral distance (LSD) in dB, for different noise types and enhancement techniques. The lower the distance, the closer is the signal to the original utterance, i.e. the better.*

filtering artifacts, such as a dampened and somewhat distorted sound in case of log-MMSE and a slight "crackling" sound in case of the SFE-WF. At very low SNRs (0dB and below) the SFE-WF enhanced sound started to have a synthetic character. Moreover, the noise seemed to get integrated into the voice of the speaker, which might be a result of keeping the noisy fine spectral structure.

Regarding the above experiments, it should be noted that we used Loizou's implementations [14] of spectral subtraction, Wiener filtering and log-spectral amplitude estimation. The required noise estimates were obtained with Martin's statistical model based voice activity detector [21], as described in more detail in [14, §11.2]. For speech enhancement with the model-based Wiener filter as well as the direct reconstruction approach from Section 3.1, we used speaker-dependent Gaussian mixture models with 128 components. With a model of this size and a frame-shift of 32, the enhancement ran in approximately 3 times real time, on a 3.00GHz Intel Xeon 5100 CPU.

5. Conclusions

We have proposed a novel, model-based Wiener filter that corrects the spectral envelope of the speech signal based on a reasonably novel speech feature enhancement approach in the log-Mel domain. Its superior performance with respect to non-stationary noise-suppression has been shown in a comparison with state-of-the-art implementations of standard speech enhancement methods.

6. References

[1] M. R. Weiss et al., "Processing speech signals to attenuate interference," *Proc. IEEE Symp. Speech Recognition*, 1974.

[2] M. W. Callahan, *Acoustic Signal Processing Based on Short-Time Spectrum*. Ph.D. dissertation, Department of Computer Science, University of Utah, Salt Lake City, Mar. 1976.

[3] J. S. Lim and A. V. Oppenheim, "All-pole modelling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 197–210, Jun. 1978.

[4] Y. Ephraim and D. Mallah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[5] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech, Audio Process.*, vol. 10, no. 6, pp. 341–351, Sep. 2002.

[6] M. Nilsson, M. Dahl, and I. Claesson, "HMM-based speech enhancement applied in non-stationary noise using cepstral features and log-nornal approximation," *Proc. DSPCS*, pp. 82–86, Dec. 2003.

[7] C. Breithaupt and R. Martin, "MMSE estimation of magnitude-squared DFT coefficients with super-Gaussian priors," *Proc. ICASSP*, pp. 896–899, Apr. 2003.

[8] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden markov models," *IEEE Trans. Audio, Speech, Signal Process.*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[9] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 4, no. 5, pp. 383–389, Sep. 1996.

[10] A. Kundu, S. Chatterjee, A. S. Murthy, and T. V. Sreenivas, "HMM-based speech enhancement applied in non-stationary noise using cepstral features and log-nornal approximation," *Proc. ICASSP*, pp. 4893–4896, Apr. 2008.

[11] Y. Shinohara and M. Akamime, "Bayesian feature enhancement using a mixture of unscented transformations for uncertainty decoding," *Proc. ICASSP*, pp. 4569–4572, Apr. 2009.

[12] F. Faubel and D. Klakow, "Estimating noise from noisy speech features with a monte carlo variant of the expectation maximization algorithm," *Proc. Interspeech*, pp. 2046–2049, Sep. 2010.

[13] F. Faubel, J. McDonough, and D. Klakow, "A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-mel domain," *Proc. Interspeech*, Sep. 2008.

[14] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, Jun. 2007.

[15] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," *Proc. ASRU*, Nov. 2005.

[16] A. 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," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[17] F. Faubel, J. McDonough, and D. Klakow, "On expectation maximization based channel and noise estimation beyond the Taylor series expansion," *Proc. ICASSP*, pp. 4294–4297, Mar. 2010.

[18] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," *Proc. ICASSP*, May 2001.

[19] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. ICASSP*, pp. 208–211, Apr. 1979.

[20] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," *Proc. ICASSP*, pp. 629–632, May 1996.

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