Spectral Noise Tracking for Improved Nonstationary Noise Robust ASR

Aleksej Chinaev[1]*, Marc Puels, Reinhold Haeb-Umbach*

Department of Communications Engineering, University of Paderborn, 33098, Paderborn, Germany
Email: {chinaev, haeb}@nt.uni-paderborn.de
Web: nt.uni-paderborn.de

Abstract

A method for nonstationary noise robust automatic speech recognition (ASR) is to first estimate the changing noise statistics and second clean up the features prior to recognition accordingly. Here, the first is accomplished by noise tracking in the spectral domain, while the second relies on Bayesian enhancement in the feature domain. In this way we take advantage of our recently proposed maximum a-posteriori based (MAP-B) noise power spectral density estimation algorithm, which is able to estimate the noise statistics even in time-frequency bins dominated by speech. We show that MAP-B noise tracking leads to an improved noise model estimate in the feature domain compared to estimating noise in speech absence periods only, if the bias resulting from the nonlinear transformation from the spectral to the feature domain is accounted for. Consequently, ASR results are improved, as is shown by experiments conducted on the Aurora IV database.

1 Introduction

ASR in the presence of nonstationary distortions is still a major challenge. This is evidenced by the interest in benchmarks on this topic, such as the series of computational hearing in multisource environments (CHiME) challenges [1]. If only a single microphone signal is available, most noise estimation algorithms rely on the assumption that noise is 'more stationary' than speech and that time-frequency bins exist, where noise can be observed alone. A multitude of noise tracking algorithms have been developed that are based on these assumptions [2]. They usually operate in the short-time Fourier transform (STFT) domain, where speech is known to have a sparse representation, such that time-frequency bins can be identified that are dominated by noise.

In [3] we have proposed a MAP-B estimator which works in the power spectral (PS) domain and can update its noise power spectral density (PSD) estimate even if speech is dominant in the time-frequency bin under consideration. The estimator computes the posterior probability density function (PDF) of the spectral noise variance in the presence of an observation, which is "distorted" by speech. Using the estimator as a postprocessor of a speech enhancement system we were able to improve the noise PSD estimates of most state-of-the-art noise trackers for a large range of noise types and signal-to-noise ratios (SNR) [4].

ASR systems, however, rely on logarithmic mel power spectral coefficients (LMPSC), mel-frequency cepstral coefficients (MFCC) or features derived from them. The tracking of noise in these domains has turned out to be quite difficult [5]. Therefore, many noise robust ASR systems rely on estimates of the noise features computed from time spans of speech absence, such as the beginning of an utterance, and which are kept constant during speech presence. However, the reliable identification of such time spans by means of a voice activity detection is a challenging task on its own right. Furthermore, such a setup is unable to follow the nonstationarity of noise during speech activity.

Several efforts have therefore been made to estimate the noise for noise robust ASR in the mel spectral domain. Dynamic noise adaptation not only models noise in the mel spectrum as a Gaussian dynamical process [6] but does the complete inference, i.e., the estimation of the clean speech posterior in this domain. In [7] it was shown that STFT-based speech enhancement can also be quite effective for ASR, if the increased variance of the noise compensated features is accounted for. Yoshioka and Nakatani argued that the preferred domain for estimating nonstationary distortions is the STFT or PS domain, while the compensation for noise for robust ASR is best done in the LMPSC or MFCC domain [8]. The parameters of the transformation of the noise PSD estimate from the PS into the LMPSC domain, which they dubbed noise model transfer (NMT) are determined in a maximum likelihood sense, employing the Expectation Maximization (EM) algorithm. They demonstrated the effectiveness of this approach for ASR in the presence of competing speakers and reverberation.

In this paper we employ the NMT approach to transform the noise PSD estimate of the MAP-B noise tracker to the MFCC domain. Rather than using the resulting noise feature vector as a point estimate, we take it as the mean of a time-variant Gaussian prior PDF of the noise. We build upon our previous work on speech feature enhancement [9], where the clean speech feature posterior PDF is computed from the mentioned Gaussian noise prior and a Gaussian mixture model (GMM) for the speech prior. We show that the use of the MAP-B noise tracker leads to an improved noise prior in the LMPSC domain, which is able to follow the time-variant noise statistics even during the presence of speech. Recognition experiments are carried out on the AURORA IV database showing the effectiveness of the proposed noise robust ASR system.

2 System Overview

Fig. 1 gives an overview of the overall system. At its core is the Bayesian feature enhancement (BFE), that operates in the MFCC domain. The output of the BFE is the cleaned-up feature vector $\hat{\mathbf{x}}_t$, estimated from the corrupted feature vector $\mathbf{y}_t$. In our approach to noise robust ASR, the time-variant a priori model of noise is constantly updated by the MAP-B noise tracking algorithm operating in the PS domain, after transforming the spectral noise PSD estimates $\hat{\sigma}^2_{N,kt}$ to the MFCC domain resulting in $\hat{\boldsymbol{\mu}}_{\mathbf{n},t}$ via the NMT approach and the following discrete cosine transform (DCT). Below we are going to describe the different components of our ASR system in more detail.

[1]This work was in part supported by DFG under grants no. Ha 3455/8 - 1 and no. Ha 3455/11 - 1.

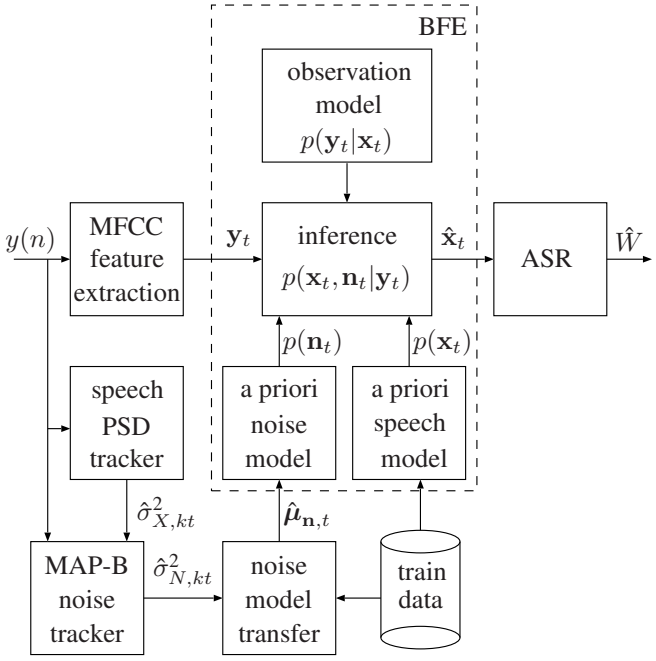

Figure 1: Block diagram of complete system

2.1 Bayesian Feature Enhancement

The goal of BFE is to compute the posterior PDF of the clean feature vector $\mathbf{x}_t$, where t denotes the frame index, given the observed noisy feature vectors $\mathbf{y}_{1:t} = \mathbf{y}_1, \ldots, \mathbf{y}_t$. When targeting noise robust speech recognition, the posterior depends on the noise feature vector $\mathbf{n}_t$. Instead of treating $\mathbf{n}_t$ as a deterministic parameter, it is modelled as a realization of a random process, to be able to deal with uncertainty. We therefore estimate the joint posterior distribution $p(\mathbf{x}_t, \mathbf{n}_t|\mathbf{y}_{1:t})$, from which the posterior $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ is obtained by marginalization, and the MMSE estimate is received via $\hat{\mathbf{x}}_t = E[\mathbf{x}_t|\mathbf{y}_{1:t}]$.

Let $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{n}_t)$. Conceptually, the posterior $p(\mathbf{z}_t|\mathbf{y}_{1:t})$ is obtained via Kalman Filter-like recursions [9], comprising the so-called prediction step:

$$p(\mathbf{z}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_{1:t-1}) \cdot p(\mathbf{z}_{t-1}|\mathbf{y}_{1:t-1}) d\mathbf{z}_{t-1} \tag{1}$$

and the following update step:

$$p(\mathbf{z}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{z}_t) \cdot p(\mathbf{z}_t|\mathbf{y}_{1:t-1}). \tag{2}$$

Here, $p(\mathbf{y}_t|\mathbf{z}_t) = p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{n}_t)$ is the observation model, which relates the observed noisy feature vector $\mathbf{y}_t$ to the underlying clean feature vector $\mathbf{x}_t$ and the noise $\mathbf{n}_t$. Further, $p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_{1:t-1}) \approx p(\mathbf{z}_t|\mathbf{z}_{t-1})$ is the a priori model. While the use of a dynamical model of speech has turned out to be crucial for the success of feature enhancement in the presence of reverberated speech [9], in the case of noise-only corruptions considered in this contribution, a static model is sufficient:

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}) \approx p(\mathbf{z}_t) = p(\mathbf{x}_t) \cdot p(\mathbf{n}_t). \tag{3}$$

While for the a priori model of speech $p(\mathbf{x}_t)$ a GMM is employed, the a priori model of noise is taken to be a single Gaussian distribution:

$$p(\mathbf{n}_t) = \mathcal{N}(\mathbf{n}_t; \boldsymbol{\mu}_{\mathbf{n},t}, \Sigma_{\mathbf{n}}). \tag{4}$$

Note, that the mean vector $\boldsymbol{\mu}_{\mathbf{n},t}$ is assumed to be time-variant, while the covariance matrix $\Sigma_{\mathbf{n}}$ is considered to be time-invariant. The time-variant mean vector $\boldsymbol{\mu}_{\mathbf{n},t}$ is obtained from the MAP-B noise tracker estimates $\hat{\sigma}^2_{N,kt}$, as will be described in the next subsection.

While environmental noise is additive in the time domain, the relationship between the MFCC feature vectors $\mathbf{y}_t, \mathbf{x}_t$ and $\mathbf{n}_t$ computed from noisy speech, speech and noise is highly nonlinear. In this work we employ the observation model

$$\mathbf{y}_t = \mathbf{D} \cdot \ln\left(e^{\mathbf{D}^{-1} \cdot \mathbf{x}_t} + e^{\mathbf{D}^{-1} \cdot \mathbf{n}_t}\right), \tag{5}$$

where $\mathbf{D}$ denotes the DCT matrix and $\mathbf{D}^{-1}$ its (pseudo) inverse. The computation of the clean speech posterior $p(\mathbf{x}_t|\mathbf{y}_t)$ relies on an iterated vector Taylor series (IVTS) approximation, where the nonlinearity in eq. (5) is linearized w.r.t. $\mathbf{x}_t$ and $\mathbf{n}_t$.

2.2 MAP-B Noise Tracker

In [3] we have presented a noise PSD estimation algorithm and its use in a single-channel speech enhancement system. Given the STFT coefficients $Y_{kt} = X_{kt} + N_{kt}$ of the noisy speech signal, where X_{kt} and N_{kt} are the STFTs of clean speech and noise signals, respectively, and where k denotes a frequency bin index, the algorithm determines an approximate MAP estimate of the spectral noise variance $\sigma^2_{N,kt} = E[|N_{kt}|^2]$. To this end the a-priori PDF $p_{\sigma^2_N}(\sigma^2)$ of the time-variant noise PSD for each frequency bin was modelled by a scaled inverse chi-squared (SICS) distribution:

$$p_{\sigma^2_N}(\sigma^2; \nu_0, \lambda^2_{kt}) = \frac{(\nu_0 \lambda^2_{kt}/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \cdot \left(\sigma^2\right)^{-\frac{\nu_0+2}{2}} \cdot e^{-\frac{\nu_0 \lambda^2_{kt}}{2\sigma^2}} \tag{6}$$

with the degrees of freedom and scale parameters ν_0 and λ^2_{kt}, respectively. (6) is a conjugate prior for the Gaussian observation PDF $p_Y(y)$ only in the case of absence of speech. In order to maintain an efficient estimation procedure in the presence of speech, the posterior $p_{\sigma^2_N|Y}(\sigma^2|y)$ was approximated by a SICS distribution with the same mode as the exact posterior PDF.

The MAP-B noise PSD estimator assumes that an estimate of the clean speech PSD $\sigma^2_{X,kt} = E[|X_{kt}|^2]$ is available. To achieve this, a speech enhancement stage is employed, which itself requires a first estimate of the noise PSD. This first coarse noise PSD estimate is afterwards refined by the MAP-B noise tracker, which acts as a postprocessor. It sholud be mentioned, that any noise tracking algorithm may be used in the first stage.

In [4] we employed eight different state-of-the-art noise PSD estimators, and the MAP-B postprocessor was able to improve the noise estimates of the majority of them under most tested noise conditions and for low to moderate SNR values. It is well known that if the SNR increases it becomes increasingly difficult to estimate the noise PSD during speech activity. The results of [10] showed that the estimation error of most noise tracking algorithms rapidly increases if the SNR is 15 dB or larger.

The MAP-B postprocessor operates in the PS domain, while the BFE and the ASR work in the MFCC domain. The noise PSD estimates must therefore be transformed to the MFCC domain. For this purpose the recently proposed NMT algorithm is employed [8].

2.3 Noise Model Transfer

The goals of the noise model transfer block in Fig. 1 are, firstly, to transform the noise PSD estimates $\hat{\sigma}^2_{N,kt}$ to the LMPSC domain, secondly, to compensate for a global bias of MAP-B estimates by means of the NMT approach [8] and, subsequently, to convert the resulting estimates to the MFCC domain. We hypothesize, that

$$\hat{\boldsymbol{\mu}}_{\mathbf{n},t} = \mathbf{D} \cdot \left(\mathrm{LM}(\hat{\boldsymbol{\sigma}}^2_{N,t}) + \mathbf{b}\right), \tag{7}$$

where $\mathrm{LM}(\cdot)$ is a shorthand notation for the processing steps of the log-mel filterbank of the ETSI standard front-end [11] i.e. the transformation from the PS domain to the LMPSC domain. Here $\hat{\boldsymbol{\sigma}}^2_{N,t} = (\hat{\sigma}^2_{N,kt}; k = 1, \ldots, K)$ is a column vector that comprises the time-variant noise PSD estimates at all K frequency bins.

The model in (7) assumes a bias vector $\mathbf{b}$, which is a result of the nonlinear logarithmic transformation and which is assumed to be constant for each utterance. In [8] a Maximum Likelihood estimate of $\mathbf{b}$ is derived in the LMPSC domain by means of the EM algorithm. While [8] employed NMT for competing speaker and reverberation estimation, it is used here for nonstationary noise tracking.

3 Experimental Results

In this section we are going to evaluate the performance of our ASR system depicted in Fig. 1 step by step. First of all the quality of noise tracking in the LMPSC domain is considered. The efficiency of the MAP-B algorithm in estimating the time-variant noise PSD has been extensively evaluated in [4]. Here, we therefore concentrate on evaluating its ability to estimate the trajectory of the noise feature vector, in conjunction with the NMT approach. Next the performance of the overall system is assessed by means of ASR results.

3.1 Aurora IV Database

Experiments are conducted on the Aurora IV database [12], where the noise signals have been artificially added to the available clean speech signals to obtain the noisy test data. This allows to compute reference noise features, which the noise estimates can be compared to. The Aurora IV database is based on the DARPA Wall Street Journal (WSJ0) Corpus. The training data are taken from the clean data of the SI-84 WSJ0 training set recorded with a Sennheiser microphone at 16 kHz sampling rate and decimated to 8 kHz. The test data consist of the 5k WSJ0 November'92 NIST evaluation test set comprising 166 utterances. 6 versions of the test set with artificially added noises at randomly chosen SNR conditions between 5 dB and 15 dB are given, namely *airport*, *babble*, *car*, *restaurant*, *street traffic* and *train station* noise.

3.2 Noise Tracking Performance

The performance of the noise tracking is evaluated in the LMPSC domain by using reference noise LMPSCs, which are calculated for each utterance by a noncausal smoothing of the known true noise LMPSC $\tilde{\mathbf{n}}_t$ over time for each mel-frequency band separately, which is realized by the Matlab function *filtfilt*(1-α, [1 -α], $\tilde{\mathbf{n}}_t$) with $\alpha = 0.95$.

For comparison purposes we start with a simple constant noise tracker in the LMPSC domain denoted by C-LMPSC. Its time-invariant estimates are calculated based on the 20 first non-speech signal frames for each utterance. This straightforward estimation procedure is often used in ASR tasks. Fig.2 shows the mean squared error (MSE) values of the C-LMPSC noise estimates averaged over time, over the LMPSC vector components and over the utterances for all 6 noise types of the Aurora IV database.

Based on the corresponding time-invariant noise PSD estimates we calculated an estimate of the clean speech PSD $\hat{\sigma}^2_{X,kt}$ by using the well known decision-directed approach, see Fig. 1 . Afterwards, we utilized an optimized MAP-B noise tracker with a degrees of freedom $\nu_0 = 80$ and a frequency dependent bias compensation factor

$$\beta_{max}(f) = \begin{cases} 2 \cdot \beta_0, & 0 < f \leq \frac{f_N}{2} \\ \left(1 + \sin^2\left(\frac{f \cdot \pi}{f_N}\right)\right) \cdot \beta_0, & \frac{f_N}{2} < f \leq f_N \end{cases} \tag{8}$$

where f_N the Nyquist frequency and $\beta_0 = 0.01$ [4]. Note that in [4] we used a frequency independent bias compensation factor. $\beta_{max}(f)$ here is heuristically chosen and motivated by the low-pass characteristic of speech PSD. The MAP-B postprocessor aims to follow the temporal changes of the nonstationary noise PSD even if speech is dominant over noise. The quality of MAP-B estimates in the LMPSC domain without any improvement through the NMT approach is shown in Fig.2.

Denoted by M-onNMT and M-offNMT we see further MSE values of the MAP-B estimates improved by the NMT approach, either in 'online' or in 'offline' mode, respectively. For estimation of the bias vector $\mathbf{b}$ in the 'offline' mode we carried out a maximum of 30 EM iterations by using the noisy features of the whole utterance. In the 'online' mode, $\mathbf{b}$ was calculated after every 20 frames based on all previous features by carrying out a maximum of 2 EM iterations on each data block. For NMT estimation we used a clean speech GMM in the LMPSC domain with 32 components and computed every component of the bias vector $\mathbf{b}$ for each dimension separately. For the purpose of comparison the MSE values of the optimal NMT approach with $\mathbf{b} = \mathbf{b}_{\text{true}}$ are also shown in Fig.2, denoted by M-optNMT.

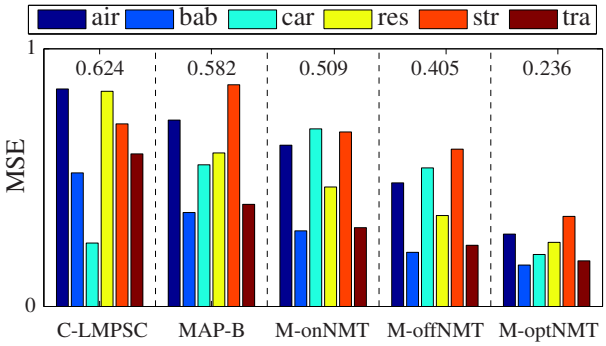

Figure 2: MSE values of different noise LMPSC estimates for the Aurora IV database.

Fig.2 shows that the MAP-B estimates are not bias free in the LMPSC domain. They absolutely need a bias correction achieved in our system by using the NMT approach. It is also striking to see that the entire system improves the noise tracking performance noticeably. By and large it can be stated that, while the MAP-B postprocessor takes care of tracking the changes of noise PSD, the NMT approach contributes decisively to reducing the residual global bias

of the MAP-B estimates in the feature domain. Moreover it should be mentioned that compared to the C-LMPSC estimates the proposed system was able to improve the tracking performance for all noise types but only for car noise, which represents a rather stationary noise type.

3.3 ASR Tests

For BFE, an a priori GMM of clean speech with 128 components and diagonal covariances is trained on MFCC feature vectors of clean training data of the Aurora IV corpus under the EM framework. Training of the Hidden Markov Models (HMMs) of the recognizer is carried out using the HMM Tool-Kit [13] on the same training data. The extracted static cepstral features are appended by dynamic features of the first and second order, giving a feature vector of length 39. For the Aurora IV task, word-internal triphone HMMs with three emitting states in a linear topology and a mixture of 10 Gaussians per state are used. The silence is modelled by a mixture of 20 Gaussians per state.

The resulting word error rates λ_{WER} for a bigram language model calculated on the enhanced features $\hat{\mathbf{x}}_t$ are given in the Tab.1 for clean speech, all 6 noise types and for different a priori noise models $p(\mathbf{n}_t)$. While the mean vectors $\hat{\boldsymbol{\mu}}_{\mathbf{n},t}$ were estimated by different noise trackers, the diagonal covariances $\hat{\boldsymbol{\Sigma}}_{\mathbf{n}}$ were always calculated based on the first 20 non-speech frames. Furthermore λ_{WER} of the Baseline (no enhancement) and the Reference (using a smoothed version of the true vector $\boldsymbol{\mu}_{\mathbf{n},t}$) are provided.

	Baseline	C-LMPSC	MAP-B	M-onNMT	M-offNMT	M-optNMT	Reference
cle	12.7	13.0	12.6	12.0	12.1	12.9	12.2
air	61.5	51.9	50.2	47.3	47.4	44.5	29.3
bab	60.6	47.0	44.9	42.6	43.0	42.0	32.0
car	39.0	19.5	18.4	17.1	16.9	17.6	15.9
res	58.8	52.7	54.0	51.9	50.5	46.9	34.8
str	58.2	43.5	45.4	43.9	42.1	41.4	30.9
tra	60.6	43.0	43.7	43.0	42.6	42.0	33.7
AVG	50.2	38.7	38.5	36.8	36.4	35.3	27.0

Table 1: Word error rates λ_{WER} on the Aurora 4 database.

It is noticeable that on average (AVG) the improved noise tracking leads to a small, however consistent, decrease of λ_{WER}. While the MAP-B estimates contribute to an improvement of λ_{WER} compared to the C-LMPSC estimator by only about 0.2 percent points, the subsequent NMT approach however reduces λ_{WER} further by up to 1.9 points in online and even 2.3 points in offline NMT mode from the maximum possible 3.4 points in the case of the M-optNMT estimates.

4 Conclusions

In this paper we have shown that the tracking of the nonstationary noise PSD in the spectral domain can lead to improved WER of an automatic speech recognizer. Spectral noise tracking was achieved by employing a MAP-based estimator. Its estimate was transformed to the feature domain using the NMT approach, which accounts for the bias introduced by the nonlinear transformation. The estimate was then used as the time-variant mean of the noise prior in Bayesian feature enhancement. While small WER gains have been achieved, there is much room for further improvement. This can be seen when comparing the results with those achievable if the true reference mean vector of the noise prior were available, see Tab.1. Further improvements are likely to be achieved by cepstral mean normalization and by nonlinearly mapping the MFCCs by a deep neural network.

References

[1] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: An overview of challenge systems and outcomes," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 162–167, Dec 2013.

[2] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4640–4643, May 2011.

[3] A. Chinaev, A. Krueger, D. H. T. Vu, and R. Haeb-Umbach, "Improved noise power spectral density tracking by a MAP-based postprocessor," in *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4041–4044, March 2012.

[4] A. Chinaev, R. Haeb-Umbach, J. Taghia, and R. Martin, "Improved single-channel nonstationary noise tracking by an optimized MAP-based postprocessor," in *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7477–7481, May 2013.

[5] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, pp. 568–580, Nov 2003.

[6] S. Rennie, P. Dognin, and P. Fousek, "Robust speech recognition using dynamic noise adaptation," in *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4592–4595, May 2011.

[7] C. Breithaupt and R. Martin, "DFT-based Speech Enhancement for Robust Automatic Speech Recognition," in *ITG Conference on Voice Communication*, pp. 1–4, Oct 2008.

[8] T. Yoshioka and T. Nakatani, "Noise Model Transfer: Novel Approach to Robustness Against Nonstationary Noise," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 2182–2192, Oct 2013.

[9] V. Leutnant, A. Krueger, and R. Haeb-Umbach, "Bayesian Feature Enhancement for Reverberation and Noise Robust Speech Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 1640–1652, Aug 2013.

[10] J. Taghia, J. Taghia, N. Mohammadiha, S. Jinqiu, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4640–4643, May 2011.

[11] ETSI ES 201 108, "Speech Processing, Transmission and Quality Aspects; Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," September 2003.

[12] H.-G. Hirsch, "Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task AU/417/02," tech. rep., STQ AURORA DSR WORKING GROUP, November 2002.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book V3.4*. Cambridge University Press, Cambridge, UK, 2006.