ON EXPECTATION MAXIMIZATION BASED CHANNEL AND NOISE ESTIMATION BEYOND THE VECTOR TAYLOR SERIES EXPANSION

Friedrich Faubel, John McDonough, Dietrich Klakow

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

ABSTRACT

In this work, we show how expectation maximization based simultaneous channel and noise estimation can be derived without a vector Taylor series expansion. The central idea is to approximate the distribution of all the random variables involved – that is noisy speech, clean speech, channel and noise – as one large, joint Gaussian distribution. Consequently, instantaneous estimates of the noise and channel distribution parameters can be obtained by conditioning the joint distribution on observed, noisy speech spectra. This approach allows for the combination of expectation maximization based channel and noise estimation with the unscented transform.

Index Terms— Speech enhancement, Speech recognition, Maximum likelihood estimation, Gaussian distributions

1. INTRODUCTION

As pointed out in 1990 by Acero [1], it is imperative to compensate both background noise and "channel effects", comprising transfer functions of microphones and room-acoustical conditions as well as differences between speakers. Some 6 years later, Moreno presented a vector Taylor series approach [2] that not only met these demands but that was also able to estimate the noise and channel parameters from noisy speech features. For the latter, the expectation maximization (EM) algorithm [3] was used, as the existence of hidden variables prevented direct maximization of the parameter likelihoods.

More recently, the vector Taylor series expansion has been replaced by the unscented transform (UT) [4, 5] – a method for approximating nonlinear transformations of Gaussian random variables that was originally developed in the context of nonlinear filtering [6]. The advantage of the unscented transform is that it is accurate at least up to the second order term of the Taylor series expansion, without the need to calculate Jacobians or Hessians. In [4], the unscented transform has been used for acoustic model compensation. In [5] it has been used for speech feature enhancement. Unfortunately, neither of the two approaches has shown how the UT can be used for noise and channel parameter estimation. In order to do just that, we derive the EM algorithm for simultaneous channel and noise estimation in a general fashion, that is, without making use of the Taylor series expansion [7, 8]. The derivation is presented in Section 4. It relies on a particular transformation, which we introduce in Section 3 in order to approximate the required, conditional distributions. The upcoming section gives an overview of the approach. Experimental results are presented in Section 5.

This work was supported by the Federal Republic of Germany through the German Research Foundation (DFG), under the research training network IRTG 715 "Language Technology and Cognitive Systems".

2. OVERVIEW OF THE APPROACH

In order to estimate the noise in a noisy utterance, Moreno [2] modeled clean speech as a Gaussian mixture random variable X with distribution

$$p_X(x) = \sum_{k=1}^{\kappa} c_k \underbrace{\mathcal{N}(x; \mu_{X|k}, \Sigma_{X|k})}_{=p_{X|k}(x)}.$$

This distribution was used as a reference for how clean speech looks like. Its parameters – that is the prior probabilities c_k, means $\mu_{X|k}$ and covariance matrices $\Sigma_{X|k}$ – were learned from a clean speech corpus. Background noise was modeled as a single Gaussian random variable N with distribution

$$p_N(n) = \mathcal{N}(n; \mu_N, \Sigma_N),$$

whose parameters were to be estimated from the noisy utterance. Both this model and the use of the EM algorithm had been proposed previously, in [9]. Moreno [2] extended that approach by considering convolutive distortions of the speech spectra, caused by differing transfer functions of the recording equipment, room-acoustical conditions and differences between speakers. These "channel effects" where taken to have a one-point distribution. In this work, they are modeled as a Gaussian random variable H with distribution

$$p_H(h) = \mathcal{N}(h; \mu_H, \Sigma_H),$$

simply because doing so allows us to treat the channel as a stacked variable in the unscented transform. Given these models, the distribution of noisy speech can be obtained by transforming the random variables according to the interaction function

$$y_t = \underbrace{x_t + h_t + \log\left(1 + e^{n_t - x_t - h_t}\right)}_{\triangleq f(x_t, h_t, n_t)}, \tag{1}$$

which defines how clean speech x_t, noise n_t and channel h_t at time t interact to form a noisy speech observation y_t in the log-Mel domain.

2.1. Expectation Maximization

Hence, the parameters $\theta = \{\mu_N, \Sigma_N, \mu_H, \Sigma_H\}$ of the noise and channel distributions can be estimated by the effect they have on the clean speech distribution. Straightforward maximum likelihood parameter estimation, however, fails, as the speech distribution is dependent on a hidden variable, K, which identifies the Gaussian that the clean speech spectrum originated from. This prompted Rose [9]

and Moreno [2] to use the expectation maximization (EM) algorithm [3]. The EM algorithm handles the hidden variable problem by iterating between two steps, which for the case of simultaneous noise and channel estimation can be described (see derivation in Section 4) as follows:

In the first step, the channel and noise parameters from the previous iteration, $\theta^{(l)} = \{\mu_N, \Sigma_N, \mu_H, \Sigma_H\}$, are used to transform each clean speech Gaussian mode $X|k$ according to (1). The resulting, transformed random variables $Y|k$ simulate how noisy speech would look like under the estimated noise and channel parameters. Consequently, the probability that the noisy speech spectrum y_t at time t originated from the k_t-th clean speech Gaussian can be evaluated as:

$$p(k_t|y_t) = \frac{p_{Y|k_t}(y_t)}{\sum_{k'_t} p_{Y|k'_t}(y_t)}. \tag{2}$$

That is performed for all t. Then, in the second step of the EM algorithm, the parameters of the channel and noise distributions are re-estimated as $\theta^{(l+1)} = \{\hat{\mu}_N, \hat{\Sigma}_N, \hat{\mu}_H, \hat{\Sigma}_H\}$, by accumulating statistics of the instantaneous maximum likelihood estimates of channel and noise for each possible k_t, weighted with the probability $p(k_t|y_t)$ that clean speech originated from that particular Gaussian, k_t. For the noise mean and covariance this gives

$$\hat{\mu}_N = \frac{\sum_{t=1}^{\tau}\sum_{k_t=1}^{\kappa} p(k_t|y_t)\tilde{\mu}_{N_t}}{\sum_{t=1}^{\tau}\sum_{k_t=1}^{\kappa} p(k_t|y_t)}, \tag{3}$$

$$\hat{\Sigma}_N = \frac{\sum_{t=1}^{\tau}\sum_{k_t=1}^{\kappa} p(k_t|y_t)\left(\tilde{\Sigma}_{N_t} + \tilde{\mu}_{N_t}\tilde{\mu}_{N_t}^T\right)}{\sum_{t=1}^{\tau}\sum_{k_t=1}^{\kappa} p(k_t|y_t)} - \hat{\mu}_N\hat{\mu}_N^T, \tag{4}$$

where $\tilde{\mu}_{N_t}$ and $\tilde{\Sigma}_{N_t}$ denote the mean and covariance of the instantaneous noise distribution $p(n_t|y_t, k_t, \theta^{(l)})$. Similarly, the channel mean and covariance are estimated as

$$\hat{\mu}_H = \frac{\sum_{t=1}^{\tau}\sum_{k_t=1}^{\kappa} p(k_t|y_t)\tilde{\mu}_{H_t}}{\sum_{t=1}^{\tau}\sum_{k_t=1}^{\kappa} p(k_t|y_t)}, \tag{5}$$

$$\hat{\Sigma}_H = \frac{\sum_{t=1}^{\tau}\sum_{k_t=1}^{\kappa} p(k_t|y_t)\left(\tilde{\Sigma}_{H_t} + \tilde{\mu}_{H_t}\tilde{\mu}_{H_t}^T\right)}{\sum_{t=1}^{\tau}\sum_{k_t=1}^{\kappa} p(k_t|y_t)} - \hat{\mu}_H\hat{\mu}_H^T. \tag{6}$$

where $\tilde{\mu}_{H_t}$ and $\tilde{\Sigma}_{H_t}$ denote the mean and covariance of the instantaneous channel distribution $p(h_t|y_t, k_t, \theta^{(l)})$. A detailed derivation is given in Section 4.

2.2. Vector Taylor Series Expansion

A problem with the above approach consists in the fact that the acoustic distortion function (1) is nonlinear. That means the transformed random variables $Y|k$ are no longer Gaussian. Neither can their distributions be obtained in an analytic fashion. The same holds for the instantaneous channel and noise distribution $p(n_t, h_t|y_t, k_t, \theta^{(l)})$. Hence, Moreno [2] proposed to approximate the interaction function (1) by a first-order Taylor series expansion about the means of the Gaussian random variables. This "local linearization" directly lead to Gaussian approximations of the transformed distributions. Second-order correction terms for the mean and covariance where also considered in [2]. More recently, [4] and [5] started replacing the Taylor series expansion by the unscented transform (UT) [6]. In [4], the UT was used to adapt the acoustic models of the recognizer to background noise (model compensation). In [5], it was used for adapting a small, auxiliary model, which was then used for speech feature enhancement. Both papers, however, considered the noise to be known in advance or estimated it on "suspected" noise only frames at the beginnings and ends of utterances. In this work, we use the UT also for estimating the noise and channel parameters.

2.3. Speech Feature Enhancement

Given the joint distribution of clean and noisy speech, X and Y, as well as a realization of the random variable Y in form of a noisy observation y_t, the minimum mean square error estimate of clean speech can be shown to be the conditional mean [7]:

$$\hat{x}_t = \sum_{k_t=1}^{\kappa} p(k_t|y_t) \underbrace{\int x_t p(x_t|y_t, k_t)\, dx_t}_{\hat{x}_{t,k_t}}. \tag{7}$$

Assuming a joint Gaussian distribution of $(X, Y)|k$, constructed as described in Section 3, the conditional mean $\hat{x}_{t,k}$ of the k-th Gaussian distribution can be calculated in analogy to (11) and (12), as:

$$\hat{x}_{t,k_t} = \mu_{X|k_t} + \Sigma_{XY|k_t}\Sigma_{YY|k_t}^{-1}\left(y_t - \mu_Y\right). \tag{8}$$

This has been presented earlier in [5].

3. CONSTRUCTING THE REQUIRED DISTRIBUTIONS

As explained in Section 2.1, the EM approach requires transforming each clean speech Gaussian mode $X|k$ to a noisy speech distribution $Y|k$, based on the interaction function (1) as well as the noise and channel parameters. For noise and channel parameter estimation, it is further necessary to know the relation of the noisy speech variable $Y|k$ to the other variables, $X|k$, N and H. Hence, we extend the transformation to

$$\bar{X} \triangleq \begin{bmatrix} X \\ H \\ N \end{bmatrix} \overset{\bar{f}}{\longmapsto} \begin{bmatrix} Y \\ X \\ H \\ N \end{bmatrix} \triangleq \bar{Y} \quad \text{with} \quad \bar{f}\left(\begin{bmatrix} x \\ h \\ n \end{bmatrix}\right) \triangleq \begin{bmatrix} f(x,h,n) \\ x \\ h \\ n \end{bmatrix}.$$

Approximating this transformation with the unscented transform [6] yields a Gaussian approximation of $p_{\bar{Y}|k}(\bar{y})$ that is accurate at least up to the second order term of the Taylor series extension:

$$p_{\bar{Y}|k}(\bar{y}) \approx \mathcal{N}\left(\bar{y}; \mu_{\bar{Y}|k}, \Sigma_{\bar{Y}\bar{Y}|k}\right) \tag{9}$$

with mean and covariance matrix

$$\underbrace{\begin{bmatrix} \mu_{Y|k} \\ \mu_{X|k} \\ \mu_H \\ \mu_N \end{bmatrix}}_{\triangleq \mu_{\bar{Y}|k}}, \quad \underbrace{\begin{bmatrix} \Sigma_{YY|k} & \Sigma_{YX|k} & \Sigma_{YH|k} & \Sigma_{YN|k} \\ \Sigma_{XY|k} & \Sigma_{XX|k} & 0 & 0 \\ \Sigma_{HY|k} & 0 & \Sigma_{HH} & 0 \\ \Sigma_{NY|k} & 0 & 0 & \Sigma_{NN} \end{bmatrix}}_{\triangleq \Sigma_{\bar{Y}\bar{Y}|k}}.$$

Consequently, the cross-covariance matrices $\Sigma_{YH|k}$ and $\Sigma_{YN|k}$ accurately capture the relation between Y and H, N up to the second order term of the Taylor series expansion. As $p_{\bar{Y}|k}(\bar{y})$ contains the distributions of Y, $(H, Y)|k$, $(N, Y)|k$ as marginal distributions,

$$p_{Y|k_t}(y_t) \approx \mathcal{N}(y_t; \mu_{Y|k}, \Sigma_{YY|k}) \tag{10}$$

is available for evaluating (2). Moreover, the instantaneous noise and channel distributions, $p(n_t|k_t, y_t, \theta^{(l)})$ and $p(h_t|k_t, y_t, \theta^{(l)})$ can be obtained by conditioning the distributions of $(N, Y)|k_t$

and $(H, Y)|k_t$ on the observation y_t, respectively. The resulting conditional Gaussian distributions [10] are:

$$p(n_t|k_t, y_t, \theta^{(l)}) \approx \mathcal{N}(n_t; \tilde{\mu}_{N_t}, \tilde{\Sigma}_{N_t}), \quad (11)$$

$$p(h_t|k_t, y_t, \theta^{(l)}) \approx \mathcal{N}(h_t; \tilde{\mu}_{H_t}, \tilde{\Sigma}_{H_t}), \quad (12)$$

where the conditional means $\tilde{\mu}_{N_t}$ and $\tilde{\mu}_{H_t}$ are calculated as

$$\begin{aligned}\tilde{\mu}_{N_t} &= \mu_N + \Sigma_{NY|k_t}\Sigma^{-1}_{YY|k_t}(y_t - \mu_{Y|k_t}),\\ \tilde{\mu}_{H_t} &= \mu_H + \Sigma_{HY|k_t}\Sigma^{-1}_{YY|k_t}(y_t - \mu_{Y|k_t}),\end{aligned}$$

and where the conditional covariances $\tilde{\Sigma}_{N_t}$, $\tilde{\Sigma}_{H_t}$ are calculated as

$$\begin{aligned}\tilde{\Sigma}_{N_t} &= \Sigma_{NN} - \Sigma_{NY|k_t}\Sigma^{-1}_{YY|k_t}\Sigma_{YN|k_t},\\ \tilde{\Sigma}_{H_t} &= \Sigma_{HH} - \Sigma_{HY|k_t}\Sigma^{-1}_{YY|k_t}\Sigma_{YH|k_t}.\end{aligned}$$

These approximations are again accurate up to the second order term of the Taylor series expansion if the unscented transform is used. If the nonlinearity of the interaction function (1) is too strong for $\bar{Y}|k$ to be approximately Gaussian, the $\bar{X}|k$ can be split into several Gaussians with smaller covariances, using, for example, the adaptive level of detail approach presented in [11].

4. A GENERAL DERIVATION OF THE EM APPROACH

In this section, we give the expectation and maximization steps that are iterated by the EM Algorithm [3] in order to find the maximum likelihood parameter set. The derivation is based on the one given in [8]. It is more general though, as it does not use the Taylor series expansion. Adding the channel, H, as a hidden variable allows for a simpler derivation of the channel parameter estimate.

The expectation step consists in calculating the auxiliary function $\mathcal{Q}(\theta|\theta^{(l)})$, that is the expected value of the of the log likelihood function, $\log \mathcal{L}(\theta; y_{1:\tau}, k_{1:\tau}, n_{1:\tau}, h_{1:\tau})$, with respect to the distribution $p(k_{1:\tau}, n_{1:\tau}, h_{1:\tau}|y_{1:\tau}, \theta^{(l)})$ of the hidden variables $k_{1:\tau}$, $n_{1:\tau}$ and $h_{1:\tau}$, given the observed data $y_{1:\tau}$ as well as the current parameter estimate $\theta^{(l)}$. Assuming statistical independence of the (y_t, k_t, n_t, h_t) for $t = 1, \ldots, \tau$ and rewriting $p(k_t, n_t, h_t|y_t, \theta^{(l)})$ as $p(n_t, h_t|k_t, y_t, \theta^{(l)})p(k_t|y_t, \theta^{(l)})$, this simplifies to:

$$\mathcal{Q}(\theta|\theta^{(l)}) = \sum_{t=1}^{\tau}\sum_{k=1}^{\kappa} p(k_t|y_t, \theta^{(l)})\mathcal{Q}_{k_t,t}(\theta|\theta^{(l)}) \quad (13)$$

with $\mathcal{Q}_{k_t,t}(\theta|\theta^{(l)})$ being defined as

$$\int\int p(n_t, h_t|k_t, y_t, \theta^{(l)}) \log p(y_t, k_t, n_t, h_t|\theta) dh_t dn_t.$$

The maximization step consists in maximizing the auxiliary function $\mathcal{Q}(\theta|\theta^{(l)})$ with respect to $\theta = \{\hat{\mu}_N, \hat{\Sigma}_N, \hat{\mu}_H, \hat{\Sigma}_H\}$, which gives the parameter $\theta^{(l+1)}$ for the next iteration:

$$\theta^{(l+1)} \triangleq \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{(l)}). \quad (14)$$

The maximum is found by differentiating $\mathcal{Q}(\theta|\theta^{(l)})$ with respect to θ and then equating it to zero. Similar as in [8], we decompose $p(y_t, k_t, n_t, h_t|\theta)$ into

$$p(y_t|k_t, n_t, h_t, \theta)\underbrace{p(h_t|n_t, k_t, \theta)}_{p(h_t|\hat{\mu}_H, \hat{\Sigma}_H)}\underbrace{p(n_t|k_t, \theta)}_{p(n_t|\hat{\mu}_N, \hat{\Sigma}_N)}\underbrace{p(k_t|\theta)}_{p(k_t)}$$

and, subsequently, rewrite $\mathcal{Q}_{k_t,t}(\theta|\theta^{(l)})$ as

$$\begin{aligned}&\int\int \log p(y_t|k_t, n_t, h_t, \theta) p(n_t, h_t|k_t, y_t, \theta^{(l)}) dh_t dn_t\\ &+ \int \log p(n_t|\hat{\mu}_N, \hat{\Sigma}_N) \underbrace{\int p(n_t, h_t|k_t, y_t, \theta^{(l)}) dh_t}_{=p(n_t|k_t, y_t, \theta^{(l)})} dn_t\\ &+ \int \log p(h_t|\hat{\mu}_H, \hat{\Sigma}_H) \underbrace{\int p(n_t, h_t|k_t, y_t, \theta^{(l)}) dn_t}_{=p(h_t|k_t, y_t, \theta^{(l)})} dh_t\\ &+ \log p(k_t).\end{aligned}$$

Due to the nonlinear character of the interaction function (1), the distributions $p(n_t|k_t, y_t, \theta^{(l)})$ and $p(h_t|k_t, y_t, \theta^{(l)})$ will be non-Gaussian, in general. Nonetheless, we make the Gaussian approximation described in Section 3. Then, taking the derivative of $\mathcal{Q}_{k_t,t}(\theta|\theta^{(l)})$ with respect to $\hat{\mu}_N$ gives

$$\frac{d\mathcal{Q}_{k_t,t}(\theta|\theta^{(l)})}{d\hat{\mu}_N} = -\hat{\Sigma}_N^{-1}\left(\tilde{\mu}_{N_t} - \hat{\mu}_N\right), \quad (15)$$

where $\tilde{\mu}_{N_t}$ is the mean of $p(n_t|k_t, y_t, \theta^{(l)})$. Taking the derivative with respect to $\hat{\mu}_H$ gives

$$\frac{d\mathcal{Q}_{k_t,t}(\theta|\theta^{(l)})}{d\hat{\mu}_H} = -\hat{\Sigma}_H^{-1}\left(\tilde{\mu}_{H_t} - \hat{\mu}_H\right) \quad (16)$$

with $\tilde{\mu}_{H_t}$ being the mean of $p(h_t|k_t, y_t, \theta^{(l)})$. Now, substituting (15) and (16) into (13), we obtain the derivatives of the auxiliary function $\mathcal{Q}(\theta|\theta^{(l)})$ with respect to $\hat{\mu}_N$ and $\hat{\mu}_H$, which when equated to zero yield the equations given in (3) and (5). The corresponding covariance matrices $\hat{\Sigma}_N$ and $\hat{\Sigma}_H$, given in (4) and (6), can be obtained in a similar fashion. That derivation is more subtle though.

5. EXPERIMENTS

In order to evaluate the performance of EM-based channel and noise estimation with the unscented transform, we performed a series of automatic speech recognition experiments. These experiments were conducted on the close talking channel of speakers 16-25 of the *multi-channel Wall Street Journal audio visual* (MC-WSJ-AV) corpus [12]. The corresponding 352 utterances, consisting of approximately 40 minutes of speech, were artificially contaminated by adding noise from the NOISEX-92 [13] database at different *signal-to-noise ratios* (SNR)s. The feature extraction of our ASR system was based on *Mel frequency cepstral coefficients* (MFCC)s. After *cepstral mean subtraction* (CMS) with variance normalization, 15 consecutive MFCC features were concatenated and subsequently reduced by *linear discriminant analysis* (LDA) to obtain the final 42-dimensional feature. The decoder used in the experiments is based on the fast on-the-fly composition of weighted finite-state transducers (WFSTs), as described in [14, §8]. The triphone acoustic model was trained with 30 hours WSJ0 and 12 hours WSJCAM0 data, resulting in 1,743 fully continuous codebooks with a total of 70,308 Gaussians. The auxiliary 128 component clean speech Gaussian mixture model, used for noise and channel estimation as well as speech enhancement, was trained on the same data set. As the obtained word error rates were quite high, we further adapted the acoustic model to the MC-WSJ-AV speakers, using maximum likelihood linear regression (MLLR) [15] speaker adaptation. That, along with increasing the beam width by about 30%, reduced the word error rate for clean speech from 41.4 to 12.7 percent.

Table 1. *Word error rate under different noise conditions, for the baseline (none) and after speech feature enhancement with the unscented transform. The considered distortions, that is noise (N) and channel (C), where either known (estimation type oracle) or estimated with the EM algorithm (estimation type EM).*

esti-mation type	cons. distor-tions	environmental noise condition					
		destroyer engine			factory2		
		05dB	10dB	15dB	05dB	10dB	15dB
none	none	85.0	61.9	40.9	58.1	32.6	21.5
oracle	N	60.9	36.5	22.7	40.8	23.9	16.3
oracle	N,C	53.0	30.5	19.1	34.3	20.5	15.9
EM	N,C	68.6	41.2	25.0	44.3	24.9	17.7
PA-PF	N	76.3	50.6	30.4	51.6	28.8	18.8

Table 1 shows the word error rates (WER)s obtained in speech recognition experiments. As a baseline (none), recognition was performed on the noisy speech features, without enhancement. In the oracle experiments, the noise distribution was perfectly known – that is, learned from the true noise spectra, per utterance. For oracle joint channel and noise compensation (oracle N,C), the channel mean was estimated from clean speech; the variance was set to a negligibly small number (10^{-3}). For EM-based parameter estimation (EM), the noise distribution was initialized with suspected silence and noise frames obtained in a previous ASR pass; the channel distribution was initialized with a zero-mean radial Gaussian distribution with a variance of 0.1. Performing unscented transform based speech feature enhancement with oracle parameters reduced the WER by up to 44% relative for noise only compensation, by up to 52% for joint channel and noise compensation. Using the estimated channel and noise parameters from the fifth iteration of the EM algorithm reduced the WER by up to 38.8% relative for destroyer engine noise (at 15dB), by up to 23.8% for factory noise (at 5dB). In general, the reduction in WER was greater for relatively stationary destroyer engine noise than it was for more non-stationary factory noise.

In addition to the results obtained with speech feature enhancement based on (7), we give WERs for particle filter based noise compensation [16] as a comparison. It should be noted that the particle filter just compensates noise, not the channel. Moreover, though both approaches were comparable in speed during enhancement, the iterations of EM-based channel and noise estimation took about 13 hours to compute on 4 Intel Xeon 5100 CPUs clocked at 3.00GHz. This enormous computational expense is not due to using the unscented transform, but to calculating the conditional expectations in (11) and (12).

6. CONCLUSIONS

We have shown how the EM algorithm for simultaneous noise and channel estimation can be derived in a general fashion. For that, we transformed all the random variables involved with one joint unscented transform in order to obtain the relations of observed noisy speech spectra to the channel and noise parameters. As an alternative to the unscented transform, numerical integration techniques might be used, such as Gauss-Hermite quadrature or Monte Carlo integration. The experiments verified that unscented transform based noise and channel estimation works in principle.

7. REFERENCES

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, 1990.

[2] P. Moreno, B. Raj, and R. Stern, "A vector Taylor series approach for environment-independent speech recognition," *Proc. ICASSP*, vol. 2, pp. 733–736, May 1996.

[3] A. P. Dempster, N. M. Laird., and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, May 1977.

[4] Y. Hu and Q. Huo, "An HMM compensation approach using unscented transformations for noisy speech recognition," *Proc. ISCLP, LNAI 4272*, pp. 346–357, 2006.

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

[6] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[7] P. J. Moreno, *Speech Recognition in Noisy Environments, PhD Thesis*, Carnegie Mellon University, Pittsburgh, Pennsylvania, 1996.

[8] D. Y. Kim, C. K. Un, and N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Communication*, vol. 24, pp. 39–49, May 1998.

[9] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with applications to speaker identification in noise," *IEEE Trans. Audio Speech Lang. Process.*, vol. 2, no. 2, pp. 245–257, Apr. 1994.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics. Springer, 2006.

[11] F. Faubel and D. Klakow, "An adaptive level of detail approach to nonlinear estimation," *Proc. ICASSP*, Mar. 2010.

[12] 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.

[13] 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.

[14] M. Wölfel and J. McDonough, *Distant Speech Recognition*, John Wiley & Sons, New York, 2009.

[15] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, pp. 171–185, 1995.

[16] 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*, Sept. 2008.