BOUNDED CONDITIONAL MEAN IMPUTATION WITH GAUSSIAN MIXTURE MODELS: A RECONSTRUCTION APPROACH TO PARTLY OCCLUDED FEATURES

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 conditional mean imputation can be bounded through the use of box-truncated Gaussian distributions. That is of interest when signals or features are partly occluded by a superimposed interference, as then the noisy observation poses an upper bound. Unfortunately, the occurring integrals are not analytic. Hence an approximate solution has to be used. In the experimental section we apply the bounded approach to the reconstruction of partly occluded speech spectra and demonstrate its superiority over the unbounded case with respect to automatic speech recognition performance.

Index Terms— Signal reconstruction, Gaussian distributions, Mean square error methods, speech enhancement, speech recognition

1. INTRODUCTION

This paper concerns incomplete or missing data problems, i.e. situations in which some part of the data is available while some other part is missing. That can arise in different settings: surveys with incompletely answered questionnaires, transmissions over lossy channels, occlusion of objects in images or – the case that we give especial attention to – occlusion of signals by noise. By occlusion we mean that a part of the signal is dominated by noise so that the observed power stemming from the interference dominates the power of the occluded part of the signal. Consequently, the observation constitutes an upper bound for the occluded part. We can also posit a lower bound as the range of values that the signal assumes might be bounded below.

Occlusion and its treatment have been studied extensively in the field of automatic speech recognition, at sites such as Sheffield [1, 2, 3], Carnegie Mellon University [4] and IDIAP [5]. Early approaches [1, 2] considered the occluded portion missing, but did not consider its boundedness. In [3] the upper bound was eventually discovered and exploited to bound the marginalization technique devised in [1] for classification with missing data. Raj introduced this into the field of missing data "imputation" by deriving a bounded *maximum a-posteriori* (MAP) estimator [4]. As an analytical solution was not available, a computationally expensive, iterative scheme had to be used. In more recent work, Raj derived a bounded minimum mean squared error estimator [6], where the distribution of speech spectra was modeled as a diagonal covariance Gaussian mixture. As a consequence, each occluded component could be treated independently, as a doubly-truncated Gaussian distribution, as explained in [7]. Here, we extend that approach to the general case of full covariance Gaussians.

The paper is organized as follows: in section 2 we derive a general minimum mean squared error estimator for the missing data problem and show how mean imputation and conditional mean imputation fit into that framework. In section 3 we introduce the box-truncated Gaussian distribution along with approximations of its normalizing constant and mean, which are used in section 4 to bound the conditional mean imputation estimate. Experimental results are shown in section 5.

2. MISSING DATA IMPUTATION

The objective of missing data imputation is to estimate the missing part of the data given the observed part, exploiting the statistical relationship between the two. In the following, we first show the existence of an optimal estimator for this problem in section 2.1. Then we go on with relating it to standard mean imputation techniques, namely conditional mean imputation and mean imputation, in sections 2.2 and 2.3. Thereby, we give concrete formulas for the case where the data follows a Gaussian distribution. That is extended to Gaussian mixtures in section 2.4.

Throughout this paper the vector of missing data is denoted by $\mathbf{x}_m$, the vector of observed data by $\mathbf{x}_o$. The vector of complete data, $\mathbf{x}$, is partitioned as $\mathbf{x} = [\mathbf{x}_m^T \ \mathbf{x}_o^T]^T$ by reordering the coefficients.

2.1. The General Minimum Mean Squared Error Solution

Let $\hat{\mathbf{x}}_m = \delta(\mathbf{x}_o)$ be an estimator for $\mathbf{x}_m$ given the observed part $\mathbf{x}_o$ of the data. Then the expected *mean squared error* (MSE) introduced by using the estimate $\hat{\mathbf{x}}_m$ instead of the true $\mathbf{x}_m$ is:

$$\begin{aligned} MSE[\delta|\mathbf{x}_o] &\triangleq E\left\{\|\delta(\mathbf{x}_o) - \mathbf{x}_m\|^2 \ |\mathbf{x}_o\right\} \\ &= \int \|\delta(\mathbf{x}_o) - \mathbf{x}_m\|^2 \, p(\mathbf{x}_m|\mathbf{x}_o) d\mathbf{x}_m \end{aligned}$$

Minimizing the MSE by taking the derivative with respect to δ and equating it to zero yields the general *Minimum Mean Squared Error* (MMSE) solution:

$$\delta_{MMSE}(\mathbf{x}_o) = \int \mathbf{x}_m p(\mathbf{x}_m|\mathbf{x}_o) d\mathbf{x}_m \tag{1}$$

2.2. Conditional Mean Imputation

Use of the MMSE estimator from the previous section is usually referred to as conditional mean imputation [2] as it consists in finding the conditional mean of $\mathbf{x}_m$ given $\mathbf{x}_o$. In the following, we will assume that the joint distribution of $\mathbf{x}_m$ and $\mathbf{x}_o$ follows a multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ with

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_m \\ \mathbf{x}_o \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_m \\ \mu_o \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{m,m} & \Sigma_{m,o} \\ \Sigma_{o,m} & \Sigma_{o,o} \end{bmatrix},$$

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

where μ is the mean, Σ the covariance matrix. Then, using Schur's decomposition to factorize the covariance matrix, it can be shown [2] that $p(\mathbf{x_m}|\mathbf{x}_o)$ is the conditional Gaussian distribution,

$$p(\mathbf{x_m}|\mathbf{x}_o) = \mathcal{N}(\mathbf{x_m}; \mu_{m|o}, \Sigma_{m|o}), \tag{2}$$

with mean and covariance

$$\begin{aligned}\mu_{m|o} &= \mu_m + \Sigma_{m,o}\Sigma_{o,o}^{-1}(\mathbf{x}_o - \mu_o),\\ \Sigma_{m|o} &= \Sigma_{m,m} - \Sigma_{m,o}\Sigma_{o,o}^{-1}\Sigma_{o,m}.\end{aligned}$$

Hence, for the case of a joint multivariate Gaussian distribution, the MMSE estimator is:

$$\delta_{CMI}(\mathbf{x}_o) = \int \mathbf{x}_m \mathcal{N}(\mathbf{x}_m|\mu_{m|o}, \Sigma_{m|o}) d\mathbf{x}_m = \mu_{m|o}. \tag{3}$$

2.3. Mean Imputation

If the missing part $\mathbf{x}_m$ of the data is assumed to be statistically independent of the observed part $\mathbf{x}_o$, the conditional pdf $p(\mathbf{x}_m|\mathbf{x}_o)$ reduces to $p(\mathbf{x}_m)$. Consequently, the MMSE estimate is just the mean of the missing part – hence the name mean imputation [2]. For the case where $\mathbf{x}_m$ follows a Gaussian distribution we have:

$$\delta_{MI}(\mathbf{x}_o) = \int \mathbf{x}_m \mathcal{N}(\mathbf{x}_m; \mu_m, \Sigma_m) d\mathbf{x}_m = \mu_m. \tag{4}$$

Note that the use of joint Gaussian distributions with diagonal covariance matrix $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_D)$ – as in [6] – inevitably leads to mean imputation.

2.4. Extension to Gaussian Mixtures

As shown in [3], conditional mean imputation with a joint multivariate Gaussian distribution can easily be extended to the Gaussian mixture case

$$p(\mathbf{x}) = \sum_{k=1}^{K} c_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k) \tag{5}$$

where c_k, μ_k and Σ_k are the probability, mean and covariance of the k-th Gaussian, respecitvely. The corresponding MMSE estimator can be shown [3] to be

$$\delta_{CMI}(\mathbf{x}_o) = \sum_{k=1}^{K} p(k|\mathbf{x}_o)\delta_{CMI,k}(\mathbf{x}_o) \tag{6}$$

where $\delta_{CMI,k}(\mathbf{x}_o)$ is the conditional mean imputation estimate for the k-th Gaussian – computed according to equation (3) – and where

$$p(k|\mathbf{x}_o) = \frac{c_k p(\mathbf{x}_o|k)}{\sum_{k'=1}^{K} c_{k'} p(\mathbf{x}_o|k')}$$

is the posterior probability of the k-th Gaussian.

3. THE BOX-TRUNCATED MULTIVARIATE GAUSSIAN DISTRIBUTION

With a box-truncated multivariate Gaussian distribution we mean a multivariate Gaussian distribution that is truncated to a D-dimensional box $[L_1, U_1] \times \cdots \times [L_D, U_D]$. It can be formally defined as

$$\mathcal{N}^{[L,U]}(x; \mu, \Sigma) \triangleq \frac{1}{c^{[L,U]}}\ \mathcal{N}(x; \mu, \Sigma)|_L^U \tag{7}$$

where $c^{[L,U]}$ is the normalization constant, $\mathcal{N}(x; \mu, \Sigma)|_L^U$ is $\mathcal{N}(x; \mu, \Sigma)$ on $\prod_{d=1}^{D} [L_d, U_d]$, zero outside.

3.1. Normalizing Constant

As truncation removes all probability mass outside the D-dimensional box given by L and U, the truncated pdf does not integrate to one, unless we recompute the normalizing constant

$$c^{[L,U]} \triangleq \int_L^U \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)} d\mathbf{x}. \tag{8}$$

That, however, constitutes a problem as (8) cannot be computed analytically. Hence, we have to resort to approximations such as Genz's Monte Carlo (MC) approach [8], which is used in current Matlab versions, but which turned out to be too slow for our application. We consider two alternatives here: firstly, a diagonal covariance approximation – which is obtained by zeroing off-diagonal elements – and secondly, the approximation method derived in the following:

Let $\Sigma^{-1} = A^T A$ be the Cholesky decomposition of Σ^{-1} with A being an upper triangular matrix. Then, substituting $A(\mathbf{x} - \mu)$ by $\mathbf{z}$ in (8) yields:

$$\begin{aligned} c^{[L,U]} &\approx \int_{L'}^{U'} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}} \sqrt{|\Sigma|} d\mathbf{z} \\ &= \int_{L'}^{U'} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} d\mathbf{z}, \end{aligned} \tag{9}$$

where $L' = A^{-1}L + \mu$, $U' = A^{-1}U + \mu$ and the multiplication by the Jacobian determinant $|d\mathbf{x}/d\mathbf{z}| = \sqrt{|\Sigma|}$ is due to the change of variables. Note that after transformation the region of integration actually is a rotated and scaled box that is no longer parallel to the axes. The assumption we make here is that the rotated box can still be reasonably approximated by an axis-parallel one, which in experiments (see section 5) gave better results than the diagonal covariance matrix approximation. Now, pulling the product out of the integral in (9) we find that the normalizing constant can be approximated as

$$c^{[L,U]} \approx \prod_{i=1}^{n} \int_{L_i'}^{U_i'} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} dz_i = \prod_{i=1}^{n} \left(\mathcal{C}(U_i') - \mathcal{C}(L_i')\right), \tag{10}$$

where $\mathcal{C}$ denotes the cumulative Gaussian distribution. If the covariance matrix Σ is diagonal (10) is exact, i.e. not an approximation. The same applies to the following mean approximation.

3.2. Mean

The mean of the box-truncated multivariate Gaussian distribution is defined as

$$\mu^{[L,U]} \triangleq \frac{1}{c^{[L,U]}} \int_L^U \mathbf{x} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)} d\mathbf{x}$$

Substituting $A(\mathbf{x} - \mu)$ by $\mathbf{z}$ as in the computation of the normalization constant (section 3.1) and again making the same approximation for the bounds yields

$$\begin{aligned} \mu^{[L,U]} &\approx \frac{1}{c^{[L,U]}} \int_{L'}^{U'} (A^{-1}\mathbf{z} + \mu) \frac{1}{\sqrt{2\pi}^n} e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}} d\mathbf{z} \\ &= -A^{-1} \underbrace{\frac{1}{c^{[L,U]}} \int_{L'}^{U'} \frac{1}{\sqrt{2\pi}^n} (-\mathbf{z}) e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}} d\mathbf{z}}_{\triangleq \mathbf{m}} \\ &\quad + \mu \frac{1}{c^{[L,U]}} \underbrace{\int_{L'}^{U'} \frac{1}{\sqrt{2\pi}^n} e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}} d\mathbf{z}}_{= c^{[L,U]}} \\ &= \mu - A^{-1}\mathbf{m} \end{aligned} \tag{11}$$

We still have to find a way to compute the remaining integral in $\mathbf{m}$. In order to do that, we break $\mathbf{m}$ into its components by first writing $\mathbf{z}$ as linear combination of standard basis vectors $\mathbf{e}_i$ and then pulling the sum over the basis vectors out of the integral:

$$\begin{aligned}\mathbf{m} &= \frac{1}{c^{[L,U]}}\int_{L'}^{U'}\sum_{i=1}^{n}(-z_i)\mathbf{e}_i\frac{1}{\sqrt{2\pi}^n}e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}}d\mathbf{z}\\ &= \sum_{i=1}^{n}\mathbf{e}_i\underbrace{\frac{1}{c^{[L,U]}}\int_{L'}^{U'}(-z_i)\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z_j^2}d\mathbf{z}}_{m_i}.\end{aligned}$$

Now, making use of the fact that the integral over the product of two independent factors $f_1(x_1)$ and $f_2(x_2)$ can be calculated as the product of the integrals over the two factors,

$$\iint f_1(x_1)f_2(x_2)dx_1dx_2 = \int f_1(x_1)dx_1\int f_2(x_2)dx_2,$$

m_i can be calculated as:

$$m_i = \frac{1}{c^{[L,U]}}\underbrace{\int_{L_i'}^{U_i'}(-z_i)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z_i^2}dz_i}_{\mathcal{N}(U_i')-\mathcal{N}(L_i')}\prod_{\substack{j=1\\j\neq i}}^{n}\underbrace{\int_{L_j'}^{U_j'}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z_j^2}dz_j}_{\mathcal{C}(U_j')-\mathcal{C}(L_j')}.$$

Expanding $c^{[L,U]}$ according to equation (10) yields

$$\mathbf{m} = \left[\frac{\mathcal{N}(U_1')-\mathcal{N}(L_1')}{\mathcal{C}(U_1')-\mathcal{C}(L_1')}\cdots\frac{\mathcal{N}(U_n')-\mathcal{N}(L_n')}{\mathcal{C}(U_n')-\mathcal{C}(L_n')}\right]^T. \tag{12}$$

Note that this is consistent with the result for one-dimensional doubly truncated Gaussian distributions [7]: in the one-dimensional case we have $\Sigma = \sigma^2$, $A^{-1} = \sigma$ and hence

$$\mu^{[L,U]} = \mu - \sigma\frac{\mathcal{N}(U')-\mathcal{N}(L')}{\mathcal{C}(U')-\mathcal{C}(L')}.$$

As mentioned earlier, (10), (11) and (12) are exact for diagonal covariance matrices. So they can be used for the diagonal covariance approximation, too.

4. BOUNDED CONDITIONAL MEAN IMPUTATION

As motivated in the introduction, the occluded part $\mathbf{x}_m$ of the signal is bounded above by the observation $\mathbf{y}_m$, below by the minimal values $\mathbf{l}_m$ that the signal can assume. In this section we use that to bound the conditional mean imputation estimate.

4.1. The Gaussian Case

As established in section 2.2, the conditional distribution of $\mathbf{x}_m$ given $\mathbf{x}_o$ is $\mathcal{N}(\mathbf{x}_\mathbf{m};\mu_{m|o},\Sigma_{m|o})$ in the Gaussian case. If $\mathbf{x}_m$ is further bounded above by $\mathbf{y}_m$, below by $\mathbf{l}_m$, the conditional Gaussian distribution has to be replaced by the box-truncated conditional Gaussian distribution $\mathcal{N}^{[\mathbf{l}_m,\mathbf{y}_m]}(\mathbf{x}_m|\mu_{m|o},\Sigma_{m|o})$. Hence, in the bounded case the MMSE estimator becomes

$$\delta_{bcmi}(\mathbf{x}_o,\mathbf{y}_d) = \int \mathbf{x}_m\mathcal{N}^{[\mathbf{l}_m,\mathbf{y}_m]}(\mathbf{x}_m|\mu_{m|o},\Sigma_{m|o})d\mathbf{x}_m, \tag{13}$$

the mean of the box-truncated conditional Gaussian distribution, which can be approximated as derived in section 3.2:

$$\mu_{m|o} - A_{m|o}^{-1}\left[\frac{\mathcal{N}(y_{m,1}')-\mathcal{N}(l_{m,1}')}{\mathcal{C}(y_{m,1}')-\mathcal{C}(l_{m,1}')}\cdots\frac{\mathcal{N}(y_{m,n}')-\mathcal{N}(l_{m,n}')}{\mathcal{C}(y_{m,n}')-\mathcal{C}(l_{m,n}')}\right]^T$$

where $\mathbf{l}_m' = A_{m|o}(\mathbf{l}_m-\mu_{m|o})$, $\mathbf{y}_m' = A_{m|o}(\mathbf{y}_m-\mu_{m|o})$ and where $A_{m|o}$ is the upper triangular matrix from the Cholesky decomposition of $\Sigma_{m|o}^{-1}$.

4.2. The Gaussian Mixture Case

That can be extended to Gaussian mixture distributions by expressing $p(\mathbf{x}_m|\mathbf{x}_o,\mathbf{y}_m)$ as marginal distribution of $p(\mathbf{x}_m,k|\mathbf{x}_o,\mathbf{y}_m)$:

$$p(\mathbf{x}_m|\mathbf{x}_o,\mathbf{y}_m) = \sum_{k=1}^{K}p(\mathbf{x}_m,k|\mathbf{x}_o,\mathbf{y}_m).$$

Then, rewriting $p(\mathbf{x}_m,k|\mathbf{x}_o,\mathbf{y}_m)$ as $p(\mathbf{x}_m|\mathbf{x}_o,\mathbf{y}_m,k)p(k|\mathbf{x}_o,\mathbf{y}_m)$, with $p(\mathbf{x}_m|\mathbf{x}_o,\mathbf{y}_m,k) = \mathcal{N}^{[\mathbf{l}_m,\mathbf{y}_m]}(x_m|\mu_{m|o},\Sigma_{m|o})$, the mean of the truncated mixture can be shown to be

$$\delta_{bcmi}(\mathbf{x}_o,\mathbf{y}_m) = \sum_{k=1}^{K}p(k|\mathbf{x}_o,\mathbf{y}_m)\delta_{bcmi,k}(\mathbf{x}_o,\mathbf{y}_m) \tag{14}$$

Thereby $\delta_{bcmi,k}$ is the bounded conditional mean imputation estimate of the kth Gaussian, $p(k|\mathbf{x}_o,\mathbf{y}_m)$ is the posterior probability of the k-th Gaussian:

$$p(k|\mathbf{x}_o,\mathbf{y}_m) = \frac{c_kp(\mathbf{x}_o,\mathbf{y}_m|k)}{\sum_{k'=1}^{K}c_{k'}p(\mathbf{x}_o,\mathbf{y}_m|k')}. \tag{15}$$

In [4], it was derived that $p(\mathbf{x}_o,\mathbf{y}_m|k)$ can be evaluated as:

$$p(\mathbf{x}_o,\mathbf{y}_m|k) = \int_{\mathbf{l}_m}^{\mathbf{y}_m}p(\mathbf{x}_m,\mathbf{x}_o|k)d\mathbf{x}_m$$

That can be rewritten by expressing $p(\mathbf{x}_m,\mathbf{x}_o|k)$ as $p(\mathbf{x}_m|\mathbf{x}_o,k)\cdot p(\mathbf{x}_o|k)$:

$$p(\mathbf{x}_o,\mathbf{y}_m|k) = p(\mathbf{x}_o|k)\int_{\mathbf{l}_m}^{\mathbf{y}_m}p(\mathbf{x}_m|\mathbf{x}_o,k)d\mathbf{x}_m$$

Then, $p(\mathbf{x}_o,\mathbf{y}_m|k)$ can be calculated as

$$p(\mathbf{x}_o|\mathbf{y}_m,k) = \mathcal{N}(\mathbf{x}_o;\mu_o,\Sigma_o)\underbrace{\int_{\mathbf{l}_m}^{\mathbf{y}_m}\mathcal{N}(\mathbf{x}_m;\mu_{m|o},\Sigma_{m|o})d\mathbf{x}_m}_{C^{[\mathbf{l}_m,\mathbf{y}_m]}} \tag{16}$$

where $C^{[\mathbf{l}_m,\mathbf{y}_m]}$ is exactly the normalizing constant of the box-truncated Gaussian distribution, which can be approximated as derived in section 3.1:

$$C^{[\mathbf{l}_m,\mathbf{y}_m]} \approx \prod_{d=1}^{n}\left(\mathcal{C}(y_{m,d}')-\mathcal{C}(l_{m,d}')\right) \tag{17}$$

with $\mathbf{l}_m'$ and $\mathbf{y}_m'$ being defined as defined earlier in this section, in the paragraph following equation (13).

5. EXPERIMENTS

In order to evaluate the performance of the bounded conditional mean imputation algorithm devised in section 4 we conducted a series of *automatic speech recognition* (ASR) experiments. Thereby speech came from the MC-WSJ-AV corpus [9], noise was added from the NOISEX-92 [10] database. As both clean speech and noise were known, we could perfectly say when noise dominated the speech spectrum and, hence, run oracle experiments to compare reconstruction methods under ideal conditions. In a second set of

Table 1. *Word error rate* (WER) for oracle experiments with destroyer engine, leopard (tank) and factory2 noise at different SNRs.

noise	SNR	reconstruction method none	CMI	DBMI	BCMI[1]	BCMI[2]
engine	05 dB	91.7	85.2	81.3	85.0	72.2
	10 dB	81.0	73.6	68.7	69.2	61.0
	15 dB	70.3	64.9	57.8	54.8	49.3
factory	05 dB	75.7	72.2	62.2	69.9	59.6
	10 dB	63.7	57.2	52.2	57.0	53.1
	15 dB	55.2	49.4	51.7	49.1	48.2
tank	05 dB	58.7	54.5	50.0	57.8	50.4
	10 dB	49.5	49.2	46.5	48.4	48.2
	15 dB	45.4	42.3	46.3	45.5	44.0

Table 2. *Word error rate*s (WER)s for particle filter experiments under different conditions.

noise	SNR	reconstruction method none	CMI	DBMI	BCMI[1]	BCMI[2]
engine	05 dB	86.7	84.3	84.6	84.4	82.5
	10 dB	73.2	71.1	70.1	69.3	67.9
	15 dB	63.3	60.0	55.4	56.3	51.5
factory	05 dB	70.6	69.9	67.0	67.5	67.7
	10 dB	53.6	53.6	51.2	51.2	51.0
	15 dB	45.8	43.7	41.6	41.5	42.4

experiment we used a particle filter to estimate which regions of the speech spectrum were occluded by noise. This gives a comparison under more realistic conditions.

In the experiments, 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 [11, §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* (GMM) used by the particle filter was trained on the same data set. Spectral reconstruction was performed in the log-Mel domain and used the same GMM, however with full covariance matrices, which were estimated in a final training pass.

The oracle experiments shown in Table 1 give a comparison of the proposed *bounded conditional mean imputation* (BCMI) method with standard *conditional mean imputation* (CMI) as well as *diagonal covariance bounded mean imputation* (DBMI) [6, 7]. Note that for CMI we enforced the upper bound by resetting those imputed components that exceeded the upper bound to the upper bound, as otherwise it consistently performed worse than the baseline. BCMI^1 uses the diagonal covariance approximation of the conditional Gaussian distribution, BCMI^2 the axis-parallel box approximation described in section 3.1. Thereby the latter outperformed the former in all the oracle experiments. It also outperformed CMI, except on tank noise at 15dB. DBMI was able to keep up with BCMI^2, however only on tank noise where both methods performed comparably. For engine noise BCMI^2 vastly outperformed all other methods, yielding a WER that was 20% lower than the baseline and up to 15% lower than that of CMI.

Table 2 shows WERs for the combination of *particle filter* based noise compensation [7] with missing feature reconstruction. Thereby the particle filter was used to both compensate the noise and simultaneously estimate the probability of occlusion as described in [7]. In the paper at hand we quantized the probability of occlusion to $\{0, 1\}$ as we do not use a soft-decision approach. While on engine noise BCMI^2 outperformed all the other methods, on factory noise it performed comparably to BCMI^1 and DBMI. An interesting result is that for factory noise at 15dB PF based noise compensation without reconstruction (none) outperformed oracle based BCMI^2. With reconstruction it improved even further. We regard that as further evidence that MMSE noise compensation and missing feature reconstruction complement each other.

6. ACKNOWLEDGMENT

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". We would also like to thank Hugo van Hamme for valuable comments on an early version of this paper.

7. REFERENCES

[1] M. Cooke, P. Green, and M. Crawford, "Handling missing data in speech recognition," *Proc. ICSLP*, Sept. 1994.

[2] A. C. Morris, M. P. Cooke, and P.D. Green, "Some solutions to the missing feature problem in data classification, with application to noise robust ASR," *Proc. ICASSP*, May 1998.

[3] L. Josifovski, M. Cooke, P. Green, and A. Vizinho, "State based imputation of missing data for robust speech recognition and speech enhancement," *Proc. Eurospeech*, pp. 2837–2840, Sept. 1999.

[4] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, 2004.

[5] P. Renevey, *Speech Recognition in Noisy Conditions Using Missing Feature Approach*, EPFL Lausanne, Lausanne, Switzerland, 2000.

[6] B. Raj and R. Singh, "Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition," *Proc. ASRU*, Nov. 2005.

[7] F. Faubel, H. Raja, J. McDonough, and D. Klakow, "Particle filter based soft-mask estimation for missing feature reconstruction," *Proc. IWAENC*, Sept. 2008.

[8] A. Genz, "Numerical computation of multivariate normal probabilities," *Journal of Computational and Graphical Statistics*, vol. 1, no. 2, pp. 141–149, June 1992.

[9] 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.

[10] 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.

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