A Comparative Study of Missing Feature Imputation Techniques

Michael Braun, Friedrich Faubel[1]*, Dietrich Klakow*

Spoken Language Systems, Saarland University, 66123 Saarbrücken, Germany
Email: {michael.braun,friedrich.faubel}@lsv.uni-saarland.de
Web: www.lsv.uni-saarland.de

Abstract

This study presents a performance comparison of different missing feature imputation techniques under ideal as well as realistic conditions. The particular focus is on recent techniques such as Raj's soft-decision bounded mean imputation approach and Gemmeke's sparse imputation. In addition to experiments with oracle masks, we evaluate the usefulness of a number of different mask estimation algorithm. This includes the neg-energy criterion and a soft version of the Max-VQ algorithm. As we gradually proceed from ideal to realistic conditions, we can investigate the sensitivity of the methods towards mismatches in the acoustic conditions as well as to errors in the mask estimates.

1 Introduction

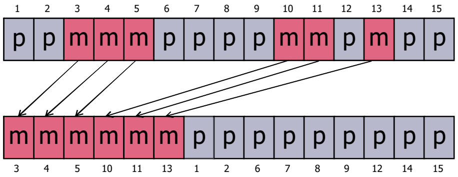
Early approaches to the treatment of speech recognition in noise go back to the mid-1970th. Around this time, Klatt [1] observed that noise actually tends to mask clean speech (log-spectral) bins when the power of noise is stronger than that of speech (in these bins). Conversely, he found that the speech spectrum is barely affected when speech is stronger. Based on these observations, Klatt proposed to simply "ignore masked bins when calculating log spectral distance scores", which would later be formulated more concisely [2] as "marginalizing over those portions which are masked by noise". Holmes and Segewick [2] even went one step further and used the observed noisy speech spectrum as an upper bound to the masked bins. This essentially laid the groundwork for *bounded marginalization*, which would later be investigated in more detail by Josivoski et al. [3]. As an alternative to the *classifier compensation methods* in [1, 2, 3], the masked parts of the speech spectrum can also be reconstructed before the classification is performed. This is what missing feature imputation (MFI) techniques [4] do in principle. Based on the type of prior model used for imputing the missing (i.e. masked) parts of a feature vector, MFI techniques may be subdivided into two categories: *statistical techniques*, which model the prior distribution of speech features as a Gaussian mixture [4, 5, 6, 7, 8, 9]; and *linear combination techniques*, which utilize a collection of exemplars [10, 11].

In this work, we present a comparative study of these techniques, especially regarding their performance under ideal and realistic conditions. For the ideal scenario, it is assumed that we are in possession of "an oracle mask generator" that tells us exactly which bins are masked by noise. This is a standard procedure for comparing MFI techniques under theoretical optimum conditions. In the realistic scenario, we investigate how well the methods perform when the masks are estimated with a real mask estimation algorithm. This tells us how well they do in practice. The particular MFI techniques which are compared in this study comprise conditional mean imputation [3], bounded conditional mean imputation [9], soft-decision bounded mean imputation [7, 8], the sparse imputation approach by Gemmeke and Cranen [10, 11] as well as a very simple approximation thereof. In addition to these MFI techniques, we compare a number of recent mask estimation algorithms, including the neg-energy criterion [4] as well as Raj's hard and soft max-VQ approaches [6, 7]. The results are presented both with and without CMLLR feature space adaptation [12].

The remaining part of the paper is organized as follows. In Section 2, we give an overview of the considered MFI techniques. Section 3 briefly explains the mask estimation algorithms that have been used. Section 4 presents the experimental results including a discussion.

2 Missing Feature Imputation

As mentioned before, we here consider two categories of missing feature imputation techniques: ones that model clean speech as a Gaussian mixture distribution and ones that model clean speech as a collection of exemplars (i.e. a set of samples). The former are described in Section 2.1. The latter are described in Section 2.2. Both approaches typically partition the feature vector $\mathbf{x}$ into a present part $\mathbf{x}_p$ as well as a missing (i.e. masked) part $\mathbf{x}_m$, which is to be imputed. Such a partitioning is easily obtained by first reordering the feature vector as portrayed below and then declaring $\mathbf{x}^T = [\mathbf{x}_m^T \;\; \mathbf{x}_p^T]$.



As the noise and therewith the masking changes in time, the reordering needs to be done individually for each frame and in dependency of the current mask $\boldsymbol{\theta} = [\theta_1 \;\; \ldots \;\; \theta_n]^T$ whose components $\theta_i \in \{0, 1\}$ identify which bins are subject to masking.

2.1 Statistical Imputation Techniques

Statistical techniques are based on modeling the distribution of clean speech as a Gaussian mixture. This means, we need to partition the means and covariance matrices in analogy to the above:

$$p(\mathbf{x}) = \sum_{k=1}^{K} c_k \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_m \\ \mathbf{x}_p \end{bmatrix}; \begin{bmatrix} \mu_{m|k} \\ \mu_{p|k} \end{bmatrix}, \begin{bmatrix} \Sigma_{mm|k} & \Sigma_{mp|k} \\ \Sigma_{pm|k} & \Sigma_{pp|k} \end{bmatrix} \right). \qquad (1)$$

In this equation, $\mathcal{N}$ denotes the Gaussian distribution, c_k denotes the weight (i.e. prior likelihood) of the k-th Gaussian; $\mu_{m|k}$ and $\mu_{p|k}$ are the missing and present part of the mean; $\Sigma_{mm|k}$ and $\Sigma_{pp|k}$ are the missing and present parts of the covariance matrix. The remaining symbols, i.e. $\Sigma_{mp|k}$ and $\Sigma_{pm|k}$, denote the cross-covariances, respectively.

2.1.1 Conditional Mean Imputation

The first MFI technique considered here is *conditional mean imputation* (CMI) [5]. It was developed in the 1990th and it essentially forms a minimum mean square error (MMSE) estimate by calculating the conditional mean of $\mathbf{x}_m$ given $\mathbf{x}_p$. This gives:

$$\hat{\mathbf{x}}_m = \sum_{k=1}^{K} c_k^+ \underbrace{\mu_m + \mathbf{\Sigma}_{mp} \mathbf{\Sigma}_{pp}^{-1} (\mathbf{x}_p - \mu_p)}_{\triangleq \mu_{m|p,k}}, \qquad (2)$$

where $\mu_{m|p,k}$ is the conditional mean of the k-th Gaussian and where c_k^+ is the probability that the observed spectrum actually

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

originated from that Gaussian:

$$c_k^+ = \frac{c_k \mathscr{N}(\mathbf{x}_p; \mu_{p|k}, \mathbf{\Sigma}_{pp|k})}{\sum_{k'=1}^K c_{k'} \mathscr{N}(\mathbf{x}_p; \mu_{p|k'}, \mathbf{\Sigma}_{pp|k'})}. \tag{3}$$

2.1.2 Bounded Conditional Mean Imputation

In more recent work, Faubel et. al [9] proposed to bound the CMI estimate through the use of a box-truncated Gaussian distribution. This approach draws on the fact that the observed noisy speech spectrum $\mathbf{y} = [\mathbf{y}_m \;\; \mathbf{y}_p]$ constitutes an upper bound for the masked part, i.e. $x_{m,i} \leq y_{m,i}$ for all i. Similarly, we can posit a lower bound at $0 \leq \mathbf{x}_{m,i}$. In order to use these bounds in the estimate from (2), each of the conditional means $\mu_{m|p,k}$ needs to be replaced by the corresponding bounded conditional mean [9]:

$$\mu_{m|p,k}^{[\mathbf{0}_m, \mathbf{y}_m]} \approx \mu_{m|p,k} - \mathbf{R}_{m|p,k}^{-1} \begin{bmatrix} \frac{\mathscr{N}(u_{k,1}) - \mathscr{N}(l_{k,1})}{\mathscr{C}(u_{l,1}) - \mathscr{C}(l_{k,1})} \\ \cdots \\ \frac{\mathscr{N}(u_{k,n}) - \mathscr{N}(l_{k,n})}{\mathscr{C}(u_{k,n}) - \mathscr{C}(l_{k,n})} \end{bmatrix}. \tag{4}$$

In this equation, $\mathbf{R}_{m|p,k}$ is the upper (right) Cholesky factor of the conditional covariance matrix $\Sigma_{m|p,k} = \Sigma_{mm|k} - \Sigma_{mp|k}\Sigma_{pp|k}^{-1}\Sigma_{pm|k}$; $\mathbf{l}_k = \mathbf{R}_{m|p,k}(\mathbf{y}_m - \mu_{m|p,k})$ and $\mathbf{u}_k = \mathbf{R}_{m|p,k}(\mathbf{0}_m - \mu_{m|p,k})$ are the transformed lower and upper bounds; and $\mathscr{C}(\cdot)$ denotes the cumulative Gaussian distribution. In addition to replacing the means, it is further necessary to replace the probabilities c_k^+ by:

$$c_k^{+[\mathbf{0}_m, \mathbf{y}_m]} = \frac{c_k \mathscr{N}(\mathbf{x}_p; \mu_{p|k}, \mathbf{\Sigma}_{pp|k}) C_k^{[\mathbf{0}_m, \mathbf{y}_m]}}{\sum_{k'=1}^K c_{k'} \mathscr{N}(\mathbf{x}_p; \mu_{p|k'}, \mathbf{\Sigma}_{pp|k'}) C_k^{[\mathbf{0}_m, \mathbf{y}_m]}}, \tag{5}$$

where $C_k^{[\mathbf{0}_m, \mathbf{y}_m]} \approx \prod_{i=1}^n \left(\mathscr{C}(u_{k,i}) - \mathscr{C}(l_{k,i})\right)$ with $\mathbf{l}_k$ and $\mathbf{u}_k$ being defined as above. Putting everything together, we finally obtain the bounded conditional mean imputation (BCMI) estimate [9]:

$$\hat{\mathbf{x}}_m = \sum_{k=1}^K c_k^{+[\mathbf{0}_m, \mathbf{y}_m]} \mu_{m|p,k}^{[\mathbf{0}_m, \mathbf{y}_m]}. \tag{6}$$

2.1.3 Soft-Decision Bounded Mean Imputation

The third MFI technique considered here is soft-decision bounded mean imputation (SD-BMI) [8]. It was originally proposed by Raj and Singh [7]. It uses diagonal covariance matrices (in contrast to full covariance matrices such as in CMI and BCMI); and it uses a soft mask $\theta_i \in [0,1]$ in order to convey the certainty that the i-th bin is masked by noise. For hard masks, i.e. $\theta_i \in \{0,1\}$, SD-BMI can be considered to be a special case of BCMI (see [9]). Due to the use of diagonal covariance matrices, the estimate can be written as interpolation of the observed noisy speech spectrum with a bounded mean imputation estimate [8]:

$$\hat{x}_i = \theta_i y_i + (1 - \theta_i) \sum_{k=1}^K p(k|\mathbf{y}, \theta) \mu_{k,i}^{[0,y_i]} \tag{7}$$

where θ is the soft mask vector, $\mathbf{y}$ is the observation vector and $\mu_{k,i}^{[0,y_i]}$ is the mean of a doubly truncated Gaussian distribution [8] with bounds 0 and y_i, i.e. the one-dimensional case of (4). In order to evaluate (7), we still need to calculate the posterior probability

$$p(k|\mathbf{y}, \theta) = \frac{c_k p(\mathbf{y}|\theta, k)}{\sum_{k'=1}^K c_{k'} p(\mathbf{y}|\theta, k')}. \tag{8}$$

Making use of the diagonality of the covariance matrix, i.e. the fact that $\Sigma_k = \mathrm{diag}([\sigma_{k,1} \;\; \ldots \;\; \sigma_{k,n}])$, the $p(\mathbf{y}|\theta, k)$ in (8) can be calculated according to [7]:

$$p(\mathbf{y}|\theta, k) = \prod_{i=1}^n p(y_i|\theta, k) \tag{9}$$

with the $p(y_i|\theta, k)$ being approximated [7, 8] as $p(y_i|\theta, k) \approx$

$$\theta_i \mathscr{N}(y_i; \mu_{k,i}, \sigma_{k,i}) + \frac{(1 - \theta_i)}{y_i} \left(\mathscr{C}(y_i; \mu_{k,i}, \sigma_{k,i}) - \mathscr{C}(0; \mu_{k,i}, \sigma_{k,i})\right).$$

2.2 Linear Combination Techniques

A different approach to impute the missing parts of a feature vector is to use *linear combination techniques*. These techniques assume that clean speech can be modeled as a collection of exemplars $\mathbf{a}_i$, $i = 1, \ldots, L$, which are randomly selected from a clean speech training corpus and then stored in an exemplar matrix $\mathbf{A}$. This matrix needs to be reordered for imputation, similar as the means of the Gaussians needed to be reordered in Section 2.1: $\mathbf{A}^T = [\mathbf{A}_m^T \;\; \mathbf{A}_p^T]$. In particular, it is assumed that the feature vector $\mathbf{x}^T = [\mathbf{x}_m^T \;\; \mathbf{x}_p^T]$ can be represented as a linear combination of the $\mathbf{a}_i$:

$$\begin{bmatrix} \mathbf{x}_m \\ \mathbf{x}_p \end{bmatrix} = \begin{bmatrix} \mathbf{A}_m \\ \mathbf{A}_p \end{bmatrix} \mathbf{w} = \sum_{i=1}^L w_i \begin{bmatrix} \mathbf{a}_{i,m} \\ \mathbf{a}_{i,p} \end{bmatrix} \tag{10}$$

with $\mathbf{w}$ specifying the contribution of each reordered exemplar. The problem of finding a suitable coefficient vector $\mathbf{w}$ can be expected to be underdetermined, as the number of exemplars is usually much larger than the dimension of the feature space, i.e. $L \gg n$ [10]. Hence, there is an abundance of possible solutions.

2.2.1 Sparse Imputation

The idea of *sparse imputation* (SI) is to select that solution $\mathbf{w}$, which has as many zero coefficients as possible, i.e. the one for which $\|\mathbf{w}\|_0$ is minimal. Unfortunately, this constrained optimization problem cannot be solved in polynomial time [11]. Hence, Gemmeke et. al [10] proposed to replace the ℓ^0-norm by an ℓ^1-norm, in which case the minimization problem can be cast as a least squares problem with an ℓ^1-penalty. In missing feature reconstruction, the coefficient vector $\mathbf{w}$ needs to be estimated from the present part $\mathbf{x}_p = \mathbf{y}_p$ of the feature vector. In this case, the minimization problem can be written [10, 11]:

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} \left(\|\mathbf{A}_p \mathbf{w} - \mathbf{y}_p\|_2 + \lambda \|\mathbf{w}\|_1\right) \tag{11}$$

with a regularization parameter λ. After finding a solution with the LASSO algorithm (with an additional non-negativity constraint on $\mathbf{w}$) [10, 11], the missing part $\mathbf{x}_m$ of the feature vector can be approximated as: $\mathbf{x}_m \approx \mathbf{A}_m \hat{\mathbf{w}}$. Finally making use of the upper bound, $\mathbf{x}_m < \mathbf{y}_m$, the estimate $\hat{\mathbf{x}}_m$ of the missing part becomes [11]:

$$\hat{\mathbf{x}}_m = \min(\mathbf{A}_m \hat{\mathbf{w}}, \mathbf{y}_m) \tag{12}$$

2.2.2 Proximative Sparse Imputation

In this work, we also use a simple approximation to the LASSO solution from [10, 11]. It consists in using that row $\mathbf{a}_j$ of A, which minimizes the least square distance between the present part $\mathbf{x}_p = \mathbf{y}_p$ of the feature vector and the present part $\mathbf{a}_{j,p}$ of the exemplar:

$$j \triangleq \underset{i}{\mathrm{argmin}} \|(\mathbf{y}_p - \mathbf{a}_{i,p})\|_2. \tag{13}$$

Then setting $\hat{\mathbf{w}} = [\hat{w}_1 \;\; \ldots \;\; \hat{w}_L]$ with $\hat{w}_i = \delta(i - j)$ (i.e. w_i is set to 1 if i equals j and to 0 otherwise), the imputation devolves to using the missing counterpart $\mathbf{a}_{j,m}$ of $\mathbf{a}_{j,p}$:

$$\hat{\mathbf{x}}_m = \min(\mathbf{A}_m \hat{\mathbf{w}}, \mathbf{y}_m) = \min(\mathbf{a}_{j,m}, \mathbf{y}_m) \tag{14}$$

We call this solution *proximative sparse imputation* (PSI).

3 Mask Estimation

In order to apply missing feature imputation in practice, we need to estimate the masks $\theta = [\theta_1 \;\; \cdots \;\; \theta_n]$ that identify which portions of the spectrum are masked by noise. The mask estimation techniques used here are all based on a Gaussian approximation of the noise distribution (in the log-Mel domain):

$$p(\mathbf{n}) = \mathscr{N}\left(\mathbf{n}; \mu^{(N)}, \Sigma^{(N)}\right) \tag{15}$$

with $\Sigma^{(N)} = \text{diag}\left([\sigma_1^{(N)} \ \cdots \ \sigma_n^{(N)}]\right)$ being a diagonal covariance matrix. This distribution is either estimated with *voice activity detection* (VAD), once for each utterance, or it is considered to be known for the particular environment (oracle noise).

3.1 Neg-Energy Criterion

The probably simplest way to obtain a binary mask $\theta_i \in \{0,1\}$, $i = 1,\dots,n$ is to use the *neg-energy criterion* (NEC) from [4]. This criterion considers a bin missing if the observed power y_i is lower than the average noise spectrum or, equivalently, if the power y_i is negative after subtraction of the noise. As a result we have the following mask estimate:

$$\hat{\theta}_i = \mathbf{1}_{\{y_i < \mu_i^{(N)}\}}(y_i) \triangleq \begin{cases} 0, & \text{if } y_i < \mu_i^{(N)} \\ 1, & \text{otherwise} \end{cases} \tag{16}$$

where $\mathbf{1}$ denotes the indicator function. This criterion was originally proposed for the Mel-domain [4]. But it can easily be translated to the log-Mel domain, as the inequality in (16) is not changed by a monotonic function such as the logarithm.

3.2 Cumulative Gauss Criterion

The *cumulative Gauss criterion* (CGC) is a mask estimation method, which we propose here as a simple extension of the neg-energy criterion. It makes use of the entire noise distribution rather than just the mean. This is achieved by calculating the expectation of $\mathbf{1}_{\{y_i<n_i\}}(y_i)$ with respect to $n_i \sim \mathcal{N}\left(\mu^{(N)}, \Sigma^{(N)}\right)$:

$$\hat{\theta}_i = \mathscr{E}\left\{\mathbf{1}_{\{y_i<n_i\}}(y_i)\right\} = \mathscr{C}\left(y_i; \mu_i^{(N)}, \sigma_i^{(N)}\right) \tag{17}$$

where $\mathscr{C}$ denotes the cumulative Gaussian density function. The result is by default a soft mask as $\hat{\theta}_i \in [0,1]$. But it can be converted to a binary mask by setting

$$\bar{\theta}_i = \begin{cases} 0, & \hat{\theta}_i < \tau \\ 1, & \text{otherwise} \end{cases} \tag{18}$$

with a threshold τ. Early experiments indicated that $\tau = 0.7$ is a reasonable value.

3.3 The Max-VQ Algorithm

The hard Max-VQ algorithm used by Raj et al. [7] models the distribution of noisy speech features $\mathbf{y}$ as a Gaussian mixture which is constructed based on the clean speech distribution $p(\mathbf{x}) = \sum_{k=1}^{K} c_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)$ from (1):

$$p(\mathbf{y}) = \sum_{k=1}^{K} c_k \mathcal{N}\left(\mathbf{y}; \mu_k^{(Y)}, \Sigma_k^{(Y)}\right). \tag{19}$$

In this equation c_k denotes the mixture weight of the k-th clean speech Gaussian and $\mu_k^{(Y)}$ and $\Sigma_k^{(Y)} = \text{diag}\left(\left[\sigma_{k,1}^{(Y)} \ \cdots \ \sigma_{k,n}^{(Y)}\right]\right)$ are calculated according to:

$$\left[\mu_{k,i}^{(Y)}, \sigma_{k,i}^{(Y)}\right] = \begin{cases} [\mu_{k,i}, \sigma_{k,i}] & \text{if } \mu_{k,i} > \mu_i^{(N)} \\ \left[\mu_i^{(N)}, \sigma_i^{(N)}\right] & \text{otherwise} \end{cases} \tag{20}$$

This means, $\mu_{k,i}^{(Y)}$ and $\sigma_{k,i}^{(Y)}$ assume the mean and variance of the k-th clean speech Gaussian if the average power of speech exceeds that of noise. They assume the parameters of the noise if the noise is stronger. From here, the max-VQ algorithm proceeds by first selecting that Gaussian component k which has the highest posterior probability

$$p(k|\mathbf{y}) = \frac{c_k \mathcal{N}\left(\mathbf{y}; \mu_k^{(Y)}, \Sigma_k^{(Y)}\right)}{\sum_{k'=1}^{K} c_{k'} \mathcal{N}\left(\mathbf{y}; \mu_{k'}^{(Y)}, \Sigma_{k'}^{(Y)}\right)};$$

and by then setting $\hat{\theta}_i$ to 0 if $\mu_{k,i} < \mu_i^{(N)}$ and to 1 otherwise. In [6], Raj and Reddy also derived a soft version of this algorithm. It calculates the mask according to:

$$\hat{\theta}_i = \sum_{k=1}^{K} p(k|\mathbf{y}) \frac{f(y_i)}{f(y_i) + g(y_i)}, \tag{21}$$

where $f(y_i) \triangleq \mathcal{N}(y_i; \mu_{k,i}, \sigma_{k,i}) \cdot \mathscr{C}\left(y_i; \mu_i^{(N)}, \sigma_i^{(N)}\right)$ and where $g(y_i) \triangleq \mathscr{C}(y_i; \mu_{k,i}, \sigma_{k,i}) \cdot \mathcal{N}\left(y_i; \mu_i^{(N)}, \sigma_i^{(N)}\right)$.

4 Experiments

This section presents the speech recognition experiments we conducted in order to compare the missing feature imputation techniques from Section 2. These experiments were performed on the multi-channel Wall Street Journal audio visual (MC-WSJ-AV) [13] corpus, more specifically: the headset data of the single speaker condition. The corresponding 352 utterances, consisting of approximately 40 minutes of speech, were artificially contaminated by adding noise from the NOISEX-92 [14] database at different signal-to-noise ratios (SNR)s. The feature extraction of the 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 was a weighted finite-state transducer (WFST) decoder. The triphone acoustic model was trained with 30 hours WSJ0 and 12 hours WSJCAM0 data. This resulted in 1,743 fully continuous codebooks with a total of 70,308 Gaussians. For statistical imputation techniques, we trained an auxiliary 128 component clean speech Gaussian mixture model on the same dataset. For linear combination techniques, we used the sliding window approach from [11] with a window length of 10. The exemplar matrix was build up from 2000 randomly selected training samples. Imputation was always performed in the log-Mel domain (with 30 bins). The resulting (imputed) spectra were fed back into the feature chain, i.e. further multiplied by a DCT matrix to obtain 20-dimensional MFCC features, and so on.

In a first experiment, we compared the missing feature imputation techniques under ideal conditions. That was achieved by using so-called "oracle" masks, which are essentially the hypothetical optimum masks that may be obtained with an ideal mask estimation algorithm (we could here calculate these masks as clean speech and noise were added under controlled conditions). Table 1 shows the resulting word error rates (WERs) for destroyer engine and factory noise, both averaged over a signal to noise ratio of 5, 10 and 15dB. In addition to first pass results (without speaker adaptation), the table gives second pass results with (unsupervised) constrained maximum likelihood linear regression (CMLLR) adaptation [12].

imputation method	no speaker adaptation		CMLLR adaptation	
	factory	destroyer	factory	destroyer
none	45.59	74.09	27.70	50.87
CMI	39.58	60.05	26.85	38.45
BCMI	25.63	35.69	17.39	28.30
SD-BMI	27.64	48.11	19.87	35.69
SI	50.95	69.07	35.00	53.22
PSI	37.18	65.66	28.04	48.59

Table 1: WERs for experiments with oracle mask

Table 1 reveals that bounded conditional mean imputation (BCMI) performs the best, with relative improvements[1] of 17.8% over soft-decision bounded mean imputation (SD-BMI), 30.0% over conditional mean imputation (CMI), 40.4% over proximate

[1] These improvements refer to the more relevant second pass and they have been averaged over destroyer engine and factory noise.

sparse imputation (PSI) and 48.2% over sparse imputation (SI). These results stand a bit in contrast to the good WERs which PSI achieved in [10, 11]. But we assumed this might be explained by a mismatch between training and testing conditions. Hence, we repeated the above experiments under ideal conditions, i.e. with an exemplar matrix that had been learned on the clean speech test data (for statistical imputation techniques we learned a new Gaussian mixture). The resulting WERs are shown in Table 2.

imputation	without speaker adapt.		with CMLLR adapt.	
method	factory	destroyer	factory	destroyer
none	45.59	74.09	27.70	50.87
CMI	42.47	66.96	26.23	44.61
BCMI	25.51	35.45	19.23	29.84
SI	30.35	42.33	25.70	34.83

Table 2: WERs after retraining the exemplar matrix (and the auxiliary Gaussian mixture) on the clean speech test set. Note that these are hypothetical results as the clean speech test data is not available in practice.

Here SI performs considerably better than CMI, which shows that linear combination techniques can indeed compete with other imputation methods. They just might be more sensitive to changes in the acoustic conditions. As a next step, we evaluated the performance of the mask estimation techniques from Section 3. This was done using both perfectly known (oracle) noise distributions as well as VAD based estimates thereof. Also note that in these experiments (as well as in all what follows), we again used the models that had been trained on the proper training set (i.e. the ones corresponding to Table 1). Figure 3 shows the resulting WERs for SD-BMI.

mask	estimated noise		oracle noise	
estimate	factory	destroyer	factory	destroyer
oracle	-	-	27.64	48.11
NEC	31.62	57.26	30.30	54.39
CGC	36.63	63.81	35.17	60.62
Max-VQ	45.64	61.96	42.98	58.32
soft Max-VQ	40.41	62.00	37.10	57.57

Table 3: WERs for SD-BMI under use of different mask estimation algorithms. Results are shown for the first pass only. Also note that we selected SD-BMI as we expected it to be the least sensitive towards mask estimation errors (as diagonal covariance matrices avert the spread of errors between dimensions).

Here, the neg-energy criterion (NEC) obviously performs the best. It is followed by the cumulative Gauss criterion (CGC), the soft Max-VQ algorithm and the binary Max-VQ algorithm. The latter two methods seemed to be a bit more sensitive towards errors in the noise estimate (larger discrepancy between estimated and oracle noise). To conclude the paper, we here finally give a direct comparison between ideal (oracle mask) and most realistic conditions (with VAD based noise estimation and NEC mask estimation). The results of these experiments are shown in Table 4. With estimated masks, SD-BMI outperforms BCMI. BCMI still performs slightly better than CMI. PSI does not really improve over the baseline. Overall, oracle masks seemed to do significantly better than estimated masks.

imputation	oracle mask		NEC mask	
method	factory	destroyer	factory	destroyer
CMI	26.45	38.45	26.94	45.67
BCMI	17.39	28.30	25.85	42.12
SD-BMI	19.87	35.69	24.79	37.63
PSI	28.04	48.59	27.71	50.42

Table 4: Oracle versus NEC masks (with VAD based noise estimation). Results are shown for the second pass only.

5 Conclusions

The comparative study conducted here reveals in particular that (a) sparse imputation techniques can be sensitive to mismatches in the acoustic conditions; (b) BCMI gives the best results with oracle masks but it is more sensitive towards mask estimation errors; (c) SD-BMI gives the best results with estimated masks and, hence, should be the method of choice in practice. The large discrepancy between oracle and estimated masks, however, also hints that the most could be gained by developing more elaborated mask estimation techniques.

References

[1] D. Klatt, "A digital filter bank for spectral matching," *Proc. ICASSP*, vol. 1, pp. 573–576, Apr. 1976.

[2] J. N. Holmes and N. C. Sedgwick, "Noise compensation for speech recognition using probabilistic models," *Proc. ICASSP*, pp. 741–744, Apr. 1986.

[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] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, "Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study," *Proc. Eurospeech*, pp. 2407 – 2410, Sept. 1999.

[5] A. Morris, M. Cooke, and P. Green, "Some solutions to the missing feature problem in data classification, with application to noise robust ASR," *Proc. ICASSP*, vol. 2, pp. 737–740, May 1998.

[6] A. M. Reddy and B. Raj, "Soft mask estimation for single channel speaker separation," *Proc. SAPA*, Oct. 2004.

[7] B. Raj and R. Singh, "Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition," *Proc. ASRU*, pp. 65–70, Nov. 2005.

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

[9] F. Faubel, J. McDonough, and D. Klakow, "Bounded conditional mean imputation with gaussian mixture models : A reconstruction approach to partly occluded features," *Proc. ICASSP*, pp. 3869–3872, Apr. 2009.

[10] J. F. Gemmeke and B. Cranen, "Using sparse representations for missing data imputation in noise robust speech recognition," *Proc. EUSIPCO*, Aug. 2008.

[11] J. F. Gemmeke and B. Cranen, "Missing data imputation using compressive sensing techniques for connected digit recognition," *Proc. Int. Conf. DSP*, pp. 1–8, July 2009.

[12] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, Jan. 1998.

[13] 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.

[14] 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.