

Missing data techniques: Feature reconstruction

Jort Florent Gemmeke¹ and Ulpu Remes²

1) Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands

2) Adaptive Informatics Research Centre, Aalto University School of Science, Finland

0.1 Introduction

Automatic speech recognition (ASR) performance degrades rapidly when speech is corrupted with increasing levels of noise. Missing data techniques (MDT) constitute a family of methods that tackle noise robust speech recognition based on the so called missing data assumption proposed in [1]. MDTs assume that (i) the noisy speech signal can be divided in speech-dominated (reliable) and noise-dominated (unreliable) spectro-temporal components prior to decoding and (ii) the unreliable elements do not retain any information about the corresponding clean speech values. This means that the clean speech values corresponding to noise-dominated components are effectively missing, and speech recognition must proceed with partially observed data.

Techniques for speech recognition with missing features divide in roughly two categories, marginalization and feature reconstruction. The marginalization approach, discussed in Chapter ??, is based on disregarding the missing components when calculating acoustic model likelihoods: likelihoods that correspond to the missing components are calculated by integrating over the full range of possible missing feature values [2, 3]. In this chapter, we focus on the reconstruction approach, where the missing values are substituted (imputed) with clean speech estimates prior to calculating the acoustic model likelihoods [4, 5, 6]. Since the reconstructed features do not contain any missing data, likelihood calculation does not need to be modified.

In general, all missing feature imputation methods employ a model of the clean speech to estimate the missing values. The models range from simple smoothness assumptions [6] to advanced statistical models and exemplar-based approaches, although the acoustic models employed by the recognizer may also be used. Given the clean speech model and a noisy observation, the missing features are estimated as the values that best match the assumptions of clean speech components at the missing locations.

Most feature reconstruction techniques are front-end based, meaning that they operate separate from the speech recognizer. Front-end imputation methods are attractive for two reasons. First, once the missing features have been replaced with clean speech estimates, any recognizer developed for clean speech can be deployed without further modifications, and second, the reconstructed features may be subjected to normalization and, for example, converted to cepstra. This is advantageous since cepstral features are known to be less correlated and better suited for processing with ASR systems based on hidden Markov model (HMM) techniques [7].

In this chapter we discuss four imputation methods in depth. First, we discuss *correlation-based imputation* and *cluster-based imputation* [5] which use statistical models to calculate clean speech estimates as the bounded maximum a posteriori (MAP) estimates. Correlation-based imputation uses a model that represents the sequence of speech frames as the output of a wide-sense stationary Gaussian process whereas in cluster-based imputation, clean speech is represented with a Gaussian mixture model (GMM). In cluster-based imputation, the bounded MAP estimates approximated as a weighted sum of estimates calculated for each Gaussian.

The third method discussed is *class-conditioned imputation*, [8, 9] in which a specific clean speech estimate is calculated for each state or Gaussian that is evaluated. Unlike the previous two methods, this requires classifier modification, and thus, imputation is no longer a strictly front-end based process. In this method, which otherwise resembles cluster-based imputation, the speech model used for imputation must have the same states and Gaussian components as the acoustic model used for speech recognition. Traditionally the acoustic model itself is used.

The final method that is treated in-depth is *sparse imputation* [10], which is an exemplar-based method. In contrast to the methods described above, speech is modeled non-parametrically as a linear combination of clean speech example spectrograms spanning multiple time-frames. Imputation is done by finding the sparsest possible linear combination of example spectrograms that accurately represents the reliable features of the noisy speech.

For each of the four methods described above, we will describe the basic concepts, discuss practical issues for implementation, and conclude with a description of possible advances proposed on the basic technique in literature. In addition to the methods described above, we will present a short overview of several other methods that can be used for reconstructing missing features in speech recognition. The methods include reconstruction based on Markov random fields [11], non-linear state-space models [11], matrix factorization [12, 13], and discrete HMMs [14].

The remainder of the chapter is organized as follows. In Section 0.2, we introduce the concept of feature reconstruction and describe the notation used in the chapter. In Sections 0.3 through 0.6, we describe the four reconstruction methods in turn, and in Section 0.7, we present short overviews of other methods. In Section 0.8, we discuss results that have been obtained with the various methods. Finally, we conclude with a discussion in Section 0.9.

0.2 Feature reconstruction

In ASR, speech is normally represented as a spectro-temporal distribution of acoustic power, a spectrogram. In noise-free conditions, the value of each time-frequency cell in this two-dimensional matrix is determined only by the speech signal. In noisy conditions, the value in

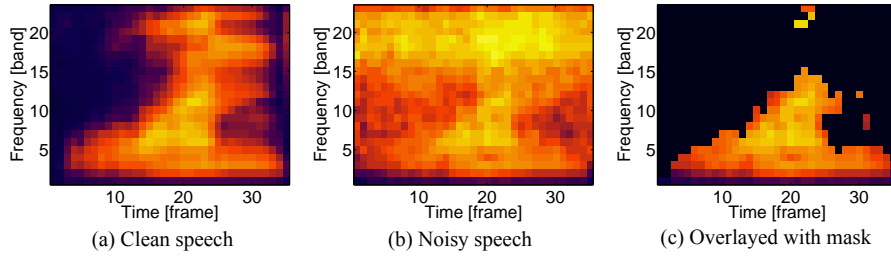


Figure 1: Figure (a) shows the spectro-temporal representation of the digit ‘one’. The horizontal axis represents time, the vertical axis represents frequency, and the intensity represents the acoustic energy. The noisy spectrogram (b) represents the clean speech after it has been artificially corrupted by suburban train noise at $\text{SNR} = -5$ dB. In Figure (c), the unreliable features of the noisy speech are marked black. We can see that a substantial part of the data needs to be reconstructed.

each cell represents a combination of speech and background noise power. Assuming noise is additive and uncorrelated with speech, the power spectrogram of noisy speech can be approximately described as the sum of the individual power spectrograms of clean speech and noise.

In ASR, the spectrographic features are often transformed to a Mel-frequency scale and compressed with a logarithmic function to mimic human hearing. Since the logarithmic compression of a two-term sum can be approximated by the logarithm of the larger of the two terms [15], it holds for noisy speech features that:

$$\mathbf{X} \approx \max(\mathbf{S}, \mathbf{N}) \quad (1)$$

with the (Mel-frequency) log-power spectrogram \mathbf{X} denoting noisy speech, \mathbf{S} denoting clean speech, and \mathbf{N} representing the background noise. The \max operator takes the element-wise maximum. Based on Equation (1), we assume that the features dominated by speech energy remain approximately uncorrupted, whereas the noise-dominated features are effectively missing.

If the clean speech and noise sources are known a priori, the observations can be divided in speech and noise dominated components. The speech-dominated, reliable features \mathbf{X}_r are directly used as estimates for the corresponding clean speech values, $\mathbf{S}_r = \mathbf{X}_r$, whereas the noise-dominated, unreliable features \mathbf{X}_u provide only an upper bound to the missing clean speech components, $\mathbf{S}_u \leq \mathbf{X}_u$. The labels that denote whether a time–frequency component $X(t, f)$, where the time index $1 \leq t \leq T$ and the frequency index $1 \leq f \leq F$, is reliable or unreliable are referred to as a missing data mask or spectrographic mask \mathbf{M} . The (binary) mask denotes reliable components as $M(t, f) = 1$ and unreliable components as $M(t, f) = 0$. See Figure 1 for an example of a noisy spectrogram and the associated missing data mask.

If the clean speech and noise sources are known a priori, a so called *oracle mask* may be constructed. In realistic situations, the location of reliable and unreliable components needs to be estimated, resulting in an *estimated mask*. Methods for mask estimation are discussed in Chapter ???. Since the labeling errors made in mask estimation can influence the

reconstruction accuracy, it may be advantageous to assess the probability that a component is reliable or unreliable, instead. The *probabilistic* or *soft mask* \mathbf{M} [16] contains continuous values between 0 and 1 that describe the probability of a feature component in \mathbf{X} being reliable.

0.3 Correlation-based imputation

0.3.1 Fundamentals

In correlation-based imputation [5], the missing values are estimated based on their statistical dependencies with reliable observations in the current and neighbouring frames. In the following sections, we introduce the clean speech model used in correlation-based imputation and derive the maximum a posteriori (MAP) estimates for the missing values based on the model. A correlation-based rule is applied to restrict the set of reliable components in order to make the estimation procedure computationally feasible.

Clean speech model

In correlation-based imputation [5], a sequence of clean speech features \mathbf{S} in the log-power domain is modelled as an output of a wide-sense stationary Gaussian process. Wide-sense stationarity means that the first and second order statistics of the feature frames $S(t)$ do not vary in time. The expected value is therefore constant, $E[S(t)] = \bar{S}$ for all t , and the covariance between any two feature frames depends only on their relative time difference, $E[S(t), S(t-l)] = \mathbf{Q}(l)$ for all t . Moreover, since the process is assumed Gaussian, the joint probability distribution of any time–frequency components $S(t, f)$ in \mathbf{S} is a Gaussian whose parameters are derived from the expected value \bar{S} and the covariance matrices $\mathbf{Q}(l)$, as discussed in Section 0.3.2.

Feature reconstruction

Given the clean speech model and a noisy spectrogram $\mathbf{X} = \mathbf{X}_r \cup \mathbf{X}_u$, estimates for the missing values \mathbf{S}_u should be chosen so that (i) the reconstructed spectrogram $\hat{\mathbf{S}} = \mathbf{S}_r \cup \hat{\mathbf{S}}_u$ fits the clean speech model while (ii) the estimates $\hat{\mathbf{S}}_u$ do not exceed the observed values \mathbf{X}_u . Clean speech estimates under the above conditions are given as

$$\hat{\mathbf{S}}_u = \operatorname{argmax}_{\xi_u} P(\xi_u | \xi_r = \mathbf{X}_r, \mathbf{S}_u \leq \mathbf{X}_u, \Lambda), \quad (2)$$

where ξ_r and ξ_u denote the random variables corresponding to \mathbf{S}_r and \mathbf{S}_u and Λ denotes the model parameters. The estimates $\hat{\mathbf{S}}_u$ are referred to as bounded MAP estimates. If the observations \mathbf{X} represent for example single words, $\hat{\mathbf{S}}_u$ can be solved using the methods discussed in Box 1. Given a longer utterance, calculations may become computationally infeasible.

Frame-based reconstruction

In order to reduce the computational load, Raj [5] proposed using a frame-based approach where the missing values $S_u(t)$ in each time frame are estimated independently. This means

that $\xi(t, f)$ and $\xi(t', k)$ that correspond to unreliable components $S_u(t, f)$ and $S_u(t', k)$ will be assumed uncorrelated when $t' \neq t$. Additionally, because correlation between any two components $\xi(t, f)$ and $\xi(t', k)$ decreases rapidly when the distance between (t, f) and (t', k) grows, many reliable components $S_r(t', k)$ do not contribute in the estimate $\hat{S}_u(t)$ and could be discarded from Equation (2) without a significant effect. The components $\xi(t', k)$ that (i) correspond to reliable observations and (ii) have a correlation greater than certain threshold α with components ξ_u that correspond to $S_u(t)$ will be denoted as ξ_n . These are the reliable components that will be used in estimating the missing values in correlation-based imputation.

To summarize, in correlation-based imputation, the random variable ξ_u will be assumed statistically independent of all feature components except ξ_n which corresponds to a subset of reliable features as discussed above. Applying the independence assumption to Equation (2),

$$\hat{S}_u(t) = \operatorname{argmax}_{\xi_u} P(\xi_u | \xi_n = X_n(t), \xi_u \leq X_u(t), \Lambda), \quad (3)$$

where $X_n(t)$ are the observations that correspond to ξ_n . Since the clean speech features \mathbf{S} are modeled as an output of a Gaussian process, the joint distribution of ξ_u and ξ_n is Gaussian, and $\hat{S}_u(t)$ can be calculated using the bounded MAP estimation methods discussed in Box 1. The estimation procedure is applied independently on each time frame to obtain the reconstructed spectrogram $\hat{\mathbf{S}} = \mathbf{S}_r \cup \hat{\mathbf{S}}_u$.

0.3.2 Implementation

Correlation-based imputation needs an estimate for the clean speech model parameters and a rule for choosing the reliable components used in reconstructing the t -th partially observed features $S(t)$. The speech data may be processed in spectrograms or in fixed-length windows centered around the current frame t . Since this does not affect estimation, we simply assume the missing values are calculated from an $F \times T$ noisy speech segment \mathbf{X} . The calculations described in the estimation section below are repeated for every time frame in the observed noisy spectrogram.

Clean speech model

All spectrograms \mathbf{S}_j in the clean speech training data are assumed independent observations of the same wide-sense stationary Gaussian process. The maximum likelihood estimate (MLE) for the feature mean is calculated as the sample average

$$\bar{S} = \frac{1}{N} \sum_j \sum_t S_j(t), \quad (4)$$

where N is the number of frames in the clean speech training data and $S_j(t)$ denotes the t -th frame of the j -th spectrogram in the training data. MLEs for the feature covariances are similarly calculated as sample covariances between the $S(t)$ and $S(t-l)$,

$$\mathbf{Q}(l) = \frac{1}{N(l)} \sum_j \sum_t (S_j(t) - \bar{S})(S_j(t-l) - \bar{S})^\top, \quad (5)$$

Box 1 Bounded estimation

Assume we are given two random variables ξ_r and ξ_u that are jointly Gaussian and an observation X so that $\xi_r = X_r$ and $\xi_u \leq X_u$, and we wish to calculate the bounded maximum a posteriori (MAP) estimate

$$\hat{S}_u(t) = \underset{\xi_u}{\operatorname{argmax}} \{P(\xi_u | \xi_r = X_r, \xi_u \leq X_u, \boldsymbol{\mu}, \boldsymbol{\Theta})\}, \quad (\text{A.1})$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Theta}$ are the mean and covariance of the joint distribution of ξ_r and ξ_u . If the covariance matrix is diagonal, the bounded MAP estimate of each $\xi(f)$ in ξ_u is the minimum of the observed upper bound $X_u(f)$ and the expected value $E[\xi(f)] = \mu(f)$. If full covariances are used, the bounded estimates cannot be solved in closed form. Since the conditional distribution is Gaussian, the bounded MAP estimates coincide with the bounded maximum likelihood (ML) and minimum mean square error (MMSE) estimates.

The bounded MAP estimation problem with Gaussian variables may be formulated as a constrained optimization task:

$$\min_{\xi} \left\{ \frac{1}{2} (\xi - \boldsymbol{\mu})^T \boldsymbol{\Theta}^{-1} (\xi - \boldsymbol{\mu}) \right\} \text{ subject to } \xi_r = X_r(t) \text{ and } \xi_u \leq X_u(t), \quad (\text{A.2})$$

where $\xi = \xi_r \cup \xi_u$. The features $\hat{\xi}$ that minimize the cost function in (A.2) maximize the joint probability distribution of ξ_r and ξ_u . Finding $\hat{\xi}$ is a general quadratic optimization problem which can be solved using iterative methods such as sequential quadratic programming (SQP). Van hamme [17] compared either using a gradient descent method to solve the optimisation problem (A.2) or using a multiplicative updates method to solve a non-negative least squares (NNLSQ) problem derived from problem (A.2). Speech recognition performance was reported to converge after 2–5 iterations with either method. In [10], the estimates were computed using a multiplicative updates method for quadratic optimization problems with non-negativity constraints [18]

Raj [5] proposed an iterative solution where the bounded MAP estimate of each $\xi(f)$ in ξ_u is calculated assuming that the other components are reliable and their values fixed in the current estimate. The estimates are initialized to the observed values, $\hat{\xi}(f) = X(f)$ for all f , and the bounded MAP estimates calculated as the minimum of $X(f)$ and the expected value $E[\xi(f) | \xi(k) = \hat{\xi}(k), k; \neq f]$ for each component in turn. The expected values are calculated as

$$E[\xi(f) | \xi(k) = \hat{\xi}(k), k; \neq f] = \mu_f + \boldsymbol{\Theta}_{fk} \boldsymbol{\Theta}_{kk}^{-1} (\hat{S}_k - \boldsymbol{\mu}_k), \quad (\text{A.3})$$

where μ_f is the expected value of $\xi(f)$ and $\boldsymbol{\Theta}_{fk}$ holds the cross-covariances between $\xi(f)$ and the other components denoted as ξ_k . The mean and covariance of the other components are denoted as $\boldsymbol{\mu}_k$ and $\boldsymbol{\Theta}_{kk}$ and their current estimates as $\hat{\xi}_k$. Iterations are continued until the estimates $\hat{\xi}(f)$ converge. It was shown in [5] that the iterative solution converges to the bounded MAP estimate of ξ_u from Equation (A.1).

where $N(l)$ is the number of frames available for estimating the l -th covariance matrix and \bar{S} is the estimated mean from Equation (4). Considering missing feature reconstruction, it is noteworthy that the mean and covariance parameters define the distribution of any subset of clean speech features in a spectrogram. More precisely, a sequence of T consecutive clean speech features concatenated into a single vector \mathbf{s} follows a Gaussian distribution with mean and covariance given as

$$\boldsymbol{\mu} = \begin{bmatrix} \bar{S} \\ \vdots \\ \bar{S} \end{bmatrix} \quad \boldsymbol{\Theta} = \begin{bmatrix} \mathbf{Q}(0) & \dots & \mathbf{Q}(T-1) \\ \vdots & & \vdots \\ \mathbf{Q}(T-1) & \dots & \mathbf{Q}(0) \end{bmatrix}, \quad (6)$$

where $\boldsymbol{\mu}$ is an FT -dimensional vector constructed by repeating the F -dimensional sample mean T times and $\boldsymbol{\Theta}$ is an $FT \times FT$ matrix constructed from the sample covariances $\mathbf{Q}(l)$.

Reliable components

In correlation-based imputation, the missing values estimates $\hat{S}_u(t)$ are calculated from the distribution of $\tilde{\xi} = \tilde{\xi}_u \cup \tilde{\xi}_n$, where ξ_u denotes the clean speech components that correspond to $S_u(t)$ and ξ_n denotes components that (i) correspond to the reliable observations $S_r(t', k)$ and (ii) according to the clean speech model, have a correlation greater than a given threshold α with at least one of the components in ξ_u . The correlation between $S(t', k)$ and $S(t, f)$ is calculated as

$$r(t - t', f, k) = \frac{q(t - t', f, k)}{\sqrt{q(0, f, f)q(0, k, k)}}, \quad (7)$$

where $q(l, f, k)$ denotes the f -th row of the k -th column of the l -th covariance matrix $\mathbf{Q}(l)$ from Equation (5).

If the speech data is processed in windows rather than full spectrograms, the window width T should be set so that the correlation between any two components more than $T/2$ frames apart does not exceed the given threshold. Raj [5] reports that the correlation between two feature components $\xi(t, f)$ and $\xi(t', k)$ falls below the proposed threshold $\alpha = 0.5$ when $|t - t'| > 5$. Note that this result depends on the frame rate and feature representation, so the minimum window width should be determined for each system separately.

Estimation

Bounded MAP estimation (Box 1) is used for computing the clean speech estimate $\hat{S}_u(t)$ based on the extended observation vector $\tilde{X}(t) = X_n(t) \cup X_u(t)$ and the mean and covariance of the corresponding clean speech distribution. To form the extended observation vector, let us reshape the $F \times T$ noisy speech segment \mathbf{X} into a single FT -dimensional vector \mathbf{x} by concatenating the subsequent time frames. Given the threshold α for choosing correlated reliable components, we can construct a binary matrix $\mathbf{U}(t)$ that extracts the components of $\tilde{X}(t)$ from \mathbf{x} , $\tilde{X}(t) = \mathbf{U}(t)\mathbf{x}$. The extended mask vector that divides $\tilde{X}(t)$ in reliable and unreliable components is similarly calculated as $\tilde{M}(t) = \mathbf{U}(t)\mathbf{m}$, where \mathbf{m} is constructed from the missing data mask \mathbf{M} . The corresponding clean speech distribution parameters are given as

$$\tilde{\boldsymbol{\mu}}(t) = \mathbf{U}(t)\boldsymbol{\mu}, \quad \tilde{\boldsymbol{\Theta}}(t) = \mathbf{U}(t)\boldsymbol{\Theta}\mathbf{U}(t)^\top, \quad (8)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Theta}$ are the mean and covariance from Equation (6). Calculating the distribution parameters is an $O(d(FT)^2)$ operation, where d denotes the number of components in $\tilde{X}(t)$.

0.4 Cluster-based imputation

0.4.1 Fundamentals

In cluster-based imputation [6], clean speech is represented with a Gaussian mixture model (GMM), and the missing values in each frame are estimated based on their statistical relationship with the reliable observations in the current frame. In the following sections, we introduce the clean speech model used in cluster-based imputation and derive the maximum a posteriori (MAP) estimates for the missing values based on the model. The estimates are approximated as a weighted sum of cluster-conditional estimates.

Clean speech model

In cluster-based imputation [6], the clean speech features $S = S(t)$ in the log-power domain are modeled as independent and identically distributed (*i.i.d.*) random variables sampled from a mixture of Gaussians:

$$P(S) = \sum_i w_i \mathcal{N}(S; \boldsymbol{\mu}_i, \boldsymbol{\Theta}_i), \quad (9)$$

where w_i is the weight of the i -th mixture component and $\boldsymbol{\mu}_i$ and $\boldsymbol{\Theta}_i$ are the mean vector and covariance matrix of the i -th component.

Feature reconstruction

Given the clean speech model and a noisy observation $X = X(t)$ with reliable and unreliable components, $X = X_r \cup X_u$, estimates for the missing values S_u in frame t should be chosen so that (i) the reconstructed feature $\hat{S} = S_r \cup \hat{S}_u$ fits the clean speech distribution model while (ii) the estimates \hat{S}_u do not to exceed the observed values X_u . Estimates for the missing values S_u under the above conditions are given as

$$\hat{S}_u = \underset{\xi_u}{\operatorname{argmax}} P(\xi_u | \xi_r = X_r, \xi_u \leq X_u, \Lambda), \quad (10)$$

where ξ_u and ξ_r denote the random variables corresponding to S_u and S_r and Λ denotes the parameters of the clean speech distribution model in Equation (9). For GMM-distributed variables, Equation (10) can be written as maximization over a weighted sum of cluster-conditional posterior probabilities,

$$\hat{S}_u = \underset{\xi_u}{\operatorname{argmax}} \left\{ \sum_i P(i|X, \Lambda) P(\xi_u | \xi_r = X_r, \xi_u \leq X_u, \Lambda_i) \right\}, \quad (11)$$

where $P(i|X, \Lambda)$ is the posterior probability for the i -th Gaussian component given the noisy observations X and $P(\xi_u | \xi_r = X_r, \xi_u \leq X_u, \Lambda_i)$ is the cluster-conditional posterior probability distribution for the unreliable features. In cluster-based imputation, Equation (11) is approximated as

$$\hat{S}_u = \sum_i P(i|X, \Lambda) \underset{\xi_u}{\operatorname{argmax}} \{ P(\xi_u | \xi_r = X_r, \xi_u \leq X_u, \Lambda_i) \}, \quad (12)$$

which is a weighted sum of cluster-conditional bounded MAP estimates for the missing values S_u . The cluster-conditional estimates can be calculated using the methods discussed in Box 1 and the posterior probabilities for clusters are calculated from the component weights and cluster-conditional observation probabilities as described below. Since the clusters are modeled with Gaussian distributions, estimates calculated from Equation (12) correspond to minimum mean square error (MMSE) estimates for S_u .

Cluster posterior probabilities

The posterior probability of the underlying clean speech feature S being associated with the i -th Gaussian component of the clean speech model is calculated as

$$P(i|X, \Lambda) = \frac{w_i P(X|\Lambda_i)}{\sum_{j=1}^I w_j P(X|\Lambda_j)}, \quad (13)$$

where $X = X_r \cup X_u$ is the noisy observation, w_i the weight of the i -th component from Equation (9), $P(X|\Lambda_i)$ the cluster-conditional observation probability, and $\Lambda_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Theta}_i\}$ denotes the mean and covariance of the i -th component. Although the clean speech model in Equation (9) is assumed to have full covariance matrices, only their diagonal components are considered when calculating the observation probabilities [6]. The probabilities are calculated as

$$P(X|\Lambda_i) = \prod_{f|S(f) \in S_r} P(\xi(f) = X_r(f)|\Lambda_i) \prod_{f|S(f) \in S_u} P(\xi(f) \leq X_u(f)|\Lambda_i), \quad (14)$$

where the likelihoods $P(\xi(f) = X_r(f)|\Lambda_i)$ correspond to the Gaussian distribution function evaluated at $X_r(f)$ and the marginal likelihoods $P(\xi(f) \leq X_u(f)|\Lambda_i)$ correspond to the Gaussian cumulative distribution function evaluated at $X_u(f)$ which is the upper bound for $S_u(f)$.

0.4.2 Implementation

In cluster-based imputation, the statistical dependencies between spectral channels $S(f)$ are modeled as a mixture of Gaussians. In the following sections, we discuss training the model with an appropriate number of components and review the feature reconstruction procedure discussed in Section 0.4.1 from a practical point of view. The calculations described in the estimation section below are repeated for every time frame in the observed noisy spectrogram.

Clean speech model

The clean speech GMM may be constructed by clustering clean speech training data in I clusters and modeling each cluster as a Gaussian. This is a simple approach that allows using any available clustering method. Alternatively, maximum likelihood estimates (MLE) for the GMM parameters can be calculated using the expectation–maximization (EM) algorithm. It alternates between calculating cluster membership probabilities for the data given the current parameter estimates (E-step) and calculating estimates for the distribution parameters

given the current membership probabilities (M-step). EM-based GMM training has been implemented in, for example, the GMMBAYES Toolbox¹ for MATLAB and the *scikits.learn* module² for Python.

The number of clusters I in the model must be such that the features in each cluster can be approximately modelled as a Gaussian. This depends on the feature representation and training data. The optimal number for missing value estimation is likely to depend on factors such as the complexity of the noisy speech recognition task and the accuracy of the missing data mask, and should be determined based on speech recognition experiments. Sometimes a small number is preferred simply because the computational complexity of cluster-based imputation grows in proportion to the number of clusters I and the performance gain from increasing the number of clusters is often small compared to the initial gain from using cluster-based imputation.

The models used in previous experiments with cluster-based imputation have varied in size and training method. A model with 512 components trained using k -means was used in [10] and a model with 5 components trained using the EM-algorithm in [19]. Since the covariances are assumed diagonal in calculating the cluster posterior probabilities, Raj et al. [6] recommend training the GMM with diagonal covariances and estimating the full covariance structure in the final pass of the EM-algorithm. A model with 128 components with full covariances estimated in the final pass was used in [20].

Estimation

Estimates for the missing clean speech features S_u are calculated from Equation (12). This corresponds to a weighted sum

$$\hat{S}_u = \sum_i \omega_i \hat{S}_u^{(i)}, \quad (15)$$

where ω_i are the cluster posterior probabilities and $\hat{S}_u^{(i)}$ the cluster-conditional estimates. The i -th cluster-conditional estimate $\hat{S}_u^{(i)}$ is calculated based on noisy observations $X = X_r \cup X_u$ and the mean and covariance parameters of the i -th Gaussian component as discussed in Box 1. The weights ω_i are calculated as a product of reliable component likelihoods and bounded marginal likelihoods associated with the unreliable components (Equation 14). The likelihoods are calculated from the Gaussian distribution function,

$$P(\xi(f) = X(f)|\Lambda_i) = \frac{1}{\sqrt{2\pi\theta_i(f)}} \exp\left(\frac{-(X(f) - \mu_i(f))^2}{2\theta_i(f)}\right) \quad (16)$$

where $\Lambda_i = \{\mu_i, \Theta_i\}$ denotes the mean and covariance of the i -th Gaussian and $\theta_i(f)$ is the f -th diagonal component of the i -th covariance matrix Θ_i . The bounded marginal likelihoods are calculated from the Gaussian cumulative distribution. They can be solved using the error function as

$$P(\xi(f) \leq X(f)|\Lambda_i) = 0.5 + 0.5 \operatorname{erf}\left(\frac{X(f) - \mu_i(f)}{\sqrt{2\theta_i(f)}}\right), \quad (17)$$

where $\operatorname{erf}(a)$ denotes the error function evaluated at a . The error function is implemented in all major programming languages.

¹Publicly available in <http://www.it.lut.fi/project/gmmbytes/>

²Publicly available in <http://scikit-learn.sourceforge.net/>

0.4.3 Advances

In cluster-based imputation, estimates for the missing values have typically been calculated as a weighted sum of cluster-conditional bounded MAP estimates (Equation 12) as proposed in [6]. These estimates correspond to the bounded minimum mean square error (MMSE) estimates calculated as

$$E[\xi_u | \xi_r = X_r, \xi_u \leq X_u, \Lambda] = \sum_i P(i|X, \Lambda_i) E[\xi_u | \xi_r = X_r, \xi_u \leq X_u, \Lambda_i], \quad (18)$$

where Λ denotes the clean speech model parameters from Equation (9), $P(i|X, \Lambda_i)$ is the posterior probability of the i -th cluster (Equation 13), and the expected value conditioned on Λ_i is the i -th cluster-conditional bounded MMSE estimate. If noise is assumed to have a uniform distribution, the cluster-conditional estimate can be solved as the expected value of a box-truncated Gaussian distribution [21]. The theoretical framework for MMSE estimation may be preferred over the MAP framework because it allows using soft masks and because a non-iterative full covariance solutions has been derived for the cluster-conditional MMSE estimates.

Soft masks

Using a soft missing data mask corresponds to assuming that each component of the observed feature X is reliable with probability $M(f)$ and unreliable with probability $1 - M(f)$. The use of soft masks in the cluster-based imputation framework was proposed in [21]. The MMSE estimate for the f -th spectral component of the underlying clean speech feature $S(f)$ is given as

$$\hat{S}(f) = M(f)X(f) + (1 - M(f))E[\xi(f) | \xi(f) \leq X(f), \Lambda]. \quad (19)$$

This may be understood as follows: If the observation $X(f)$ is reliable, the MMSE estimate for $S(f)$ is $X(f)$. This is true with probability $M(f)$. If the observations is unreliable, the estimate for $S(f)$ must be calculated from Equation (18). This is true with probability $1 - M(f)$. Note that all the components are assumed missing when calculating the estimate from Equation (18).

Estimation

Raj and Singh [21] derived a bounded MMSE estimator for Gaussian distributed variables $\xi = \xi_r \cup \xi_u$ assuming diagonal covariances. If full covariances Θ_i are used, calculating the exact solution requires iterative approaches such as discussed in Box 1. Faubel et al. [20] proposed an approximate full covariance solution calculated from the conditional distribution of ξ_u given ξ_r as

$$E[\xi_u | \xi_r = X_r, \xi_u \leq X_u, \Lambda_i] \approx \boldsymbol{\mu}_{u|r} - \mathbf{A}_{u|r}^{-1} \mathbf{p}, \quad (20)$$

where $\boldsymbol{\mu}_{u|r}$ is the conditional distribution mean, $\mathbf{A}_{u|r}$ the upper triangular matrix from the Cholesky decomposition of the inverse of the conditional distribution covariance $\Theta_{u|r}$, and \mathbf{p} is an $n \times 1$ vector whose components are a function of the conditional model parameters. The variable n denotes the number of unreliable observations in the current

frame. The conditional distribution parameters have closed-form solutions. Faubel et al. [20] also presented an approximate full covariance solution for the observation probabilities in Equation (13).

0.5 Class-conditioned imputation

0.5.1 Fundamentals

The class-conditioned imputation approaches employ the same conditional mean imputation principle as correlation and cluster-based imputation. However, instead of using a separate clean speech model, estimates for the missing values are calculated from the acoustic models, and a separate clean speech estimate is calculated for each acoustic model state [4, 8] or distribution component [22]. In the following sections, the acoustic models are assumed to use the log-power feature representation used in calculating the missing data mask and the acoustic model states are assumed to have been modeled as GMM distributions. The state and Gaussian-conditioned clean speech estimates are calculated using minimum mean square error (MMSE) estimation.

State-conditioned imputation

Front-end methods such as correlation and cluster-based imputation replace the observed features $X = X(t)$ with the reconstructed features $\hat{S} = S_r \cup \hat{S}_u$ in calculating the acoustic model likelihoods. In state-conditioned imputation [4, 8], the likelihood for each state q is calculated based on a reconstructed feature whose missing values have been estimated using the GMM distribution associated with the same state,

$$P(X|q) = \sum_i w_{iq} P(E[\xi|\xi_r = X_r, \xi_u \leq X_u, \Lambda_q]|\Lambda_{iq}), \quad (21)$$

where $X = X_r \cup X_u$ is the observed noisy feature in log-power domain and ξ_r and ξ_u denote the random variables corresponding to the reliable and unreliable clean speech components S_r and S_u . The parameters of the q -th state distribution and the i -th component of the q -th state distribution are denoted as Λ_q and Λ_{iq} , respectively, and finally, the expected value of $\xi = \xi_r \cup \xi_u$ conditioned on the observed features and state distribution parameters, $E[\xi|\xi_r = X_r, \xi_u \leq X_u, \Lambda_q]$, is the state-conditioned MMSE estimate for the clean speech features.

MMSE estimates for the reliable components S_r are the observed values X_r and the state-conditioned MMSE estimates for the missing values S_u are calculated from the q -th state distribution as

$$E[\xi_u|\xi_r = X_r, \xi_u \leq X_u, \Lambda_q] = \sum_i P(i|X, \Lambda_{iq}) E[\xi_u|\xi_r = X_r, \xi_u \leq X_u, \Lambda_{iq}], \quad (22)$$

where $P(i|X, \Lambda_{iq})$ is the posterior probability for the i -th Gaussian component given the noisy observations X . Since the acoustic model covariances are typically assumed diagonal, the Gaussian-conditioned bounded MMSE estimate $E[\xi_u|\xi_r = X_r, \xi_u \leq X_u, \Lambda_{iq}]$ can be calculated as the minimum of the observed features X_u and the unbounded MMSE estimate,

$$E[\xi_u|\xi_r = X_r, \xi_u \leq X_u, \Lambda_{iq}] = \min\{X_u, E[\xi_u|\xi_r = X_r, \Lambda_{iq}]\}. \quad (23)$$

Since the covariance matrices Θ_{iq} are assumed diagonal, the unbounded MMSE estimate is the expected value of ξ_u .

Gaussian-conditioned imputation

Van hamme [22] observed that the speech recognition performance improves if missing values are estimated specifically for each Gaussian component rather than for each state in the acoustic model. Given a noisy observation $X = X_r \cup X_u$, the state likelihoods are calculated as

$$P(X|q) = \sum_i w_{iq} P(E[\xi|\xi_r = X_r, \xi_u \leq X_u, \Lambda_{iq}]|\Lambda_{iq}), \quad (24)$$

where the Gaussian-conditioned bounded MMSE estimates $E[\xi|\xi_r = X_r, \xi_u \leq X_u, \Lambda_{iq}]$ for the reliable components are the observed values X_r , and for the missing values, the estimates can be calculated as the minimum of the observed features X_u and the unbounded MMSE estimates (Equation 23).

0.5.2 Implementation

Estimation

Gaussian-conditioned MMSE estimates for the missing values S_u may be calculated as the minimum of the observed features X_u and the expected value μ_u ,

$$E[\xi_u|\xi_r = X_r, \xi_u \leq X_u, \Lambda_{iq}] = \min\{\mathbf{X}_u, \boldsymbol{\mu}_u\}, \quad (25)$$

where μ_u denotes the components of $\boldsymbol{\mu}_{iq}$ that correspond to the missing features. The state-conditioned MMSE estimates defined in Equation (22) are calculated as a weighted sum of the Gaussian-conditioned estimates. The weights are cluster posterior probabilities calculated as a product of reliable component likelihoods and bounded marginal likelihoods associated with the unreliable components. The likelihoods are calculated from Equations (16)–(17) in Section 0.4.2.

Classifier modification

In principle, the way likelihoods are calculated does not need to be modified even when class-conditional imputation is used, but since the likelihoods are calculated based on either a state or Gaussian-dependent estimate or the reliable observation, it is necessary to implement some modifications in the likelihood calculating module of the speech recognizer being used. In practice, since modification are in any case necessary, the most computationally efficient solution may be to slightly modify the likelihood calculation. For example, for Gaussian-conditioned imputation, the likelihood calculation for unreliable features $X(f) \in \mathbf{X}_u$ can be written as

$$P(\xi(f) = X(f)) = \frac{1}{\sqrt{2\pi\theta_{iq}(f)}} \exp\left(\frac{-(\min\{X(f), \mu_{iq}(f)\} - \mu_{iq}(f))^2}{2\theta_{iq}(f)}\right). \quad (26)$$

0.5.3 Advances

Cepstral domain imputation

Features in the log-spectral domain are not attractive for speech recognition because they tend to be correlated, and hence can only be modeled effectively using full covariances. In automatic speech recognition, a decorrelating linear transformation such as the discrete cosine transformation (DCT) is used on the log-spectra. Class-conditioned imputation is not a front-end based feature reconstruction method, and thus it is not as straightforward to transform the estimates to the cepstral domain prior to evaluating the class-conditional clean speech estimate.

For Gaussian-dependent imputation, a method for doing imputation in the cepstral domain was proposed in [22]. Given a linear transformation \mathbf{A} , decorrelated features are calculated as

$$\mathbf{s}(t) = \mathbf{A}S(t), \quad (27)$$

where \mathbf{A} is the DCT-matrix. The MMSE estimate for cepstral features $\mathbf{s} = \mathbf{s}(t)$ is calculated from the distribution of cepstral features as

$$\hat{\mathbf{s}} = \underset{\xi}{\operatorname{argmax}} P(\mathbf{A}\xi | \xi_r = \mathbf{X}_r, \xi_u \leq \mathbf{X}_u, \lambda_{iq}), \quad (28)$$

where $\xi = \xi_r \cup \xi_u$ is the random variable corresponding to spectral clean speech features and $X = X(t)$ is the observed noisy speech feature in log-spectral domain. The model parameters $\lambda_{iq} = \{\boldsymbol{\mu}_{iq}, \boldsymbol{\Theta}_{iq}\}$ define the distribution of the cepstral features \mathbf{s} in the i -th Gaussian of the q -th acoustic model state. Equation (28) may be formulated as constrained optimization task:

$$\underset{\xi}{\operatorname{argmin}} \left\{ \frac{1}{2} (\mathbf{A}\xi - \boldsymbol{\mu}_{iq})^\top \boldsymbol{\Theta}_{iq}^{-1} (\mathbf{A}\xi - \boldsymbol{\mu}_{iq}) \right\} \text{ subject to } \xi_r = \mathbf{X}_r \text{ and } \xi_u \leq \mathbf{X}_u. \quad (29)$$

The clean speech features \hat{S} that minimize the cost function in (29) maximize the probability in Equation (28). Moreover, the optimization problem (29) may be written as optimization in the spectral domain,

$$\underset{\xi}{\operatorname{argmin}} \left\{ \frac{1}{2} (\xi - \boldsymbol{\mu}_S)^\top \mathbf{P} (\xi - \boldsymbol{\mu}_S) \right\} \text{ subject to } \xi_r = \mathbf{X}_r \text{ and } \xi_u \leq \mathbf{X}_u, \quad (30)$$

where $\boldsymbol{\mu}_S$ is the log-spectral domain mean that corresponds to the cepstral mean $\boldsymbol{\mu}_{iq}$ and the precision matrix \mathbf{P} is constructed as

$$\mathbf{P} = \mathbf{A}^\top \boldsymbol{\Theta}_C^{-1} \mathbf{A} + \lambda \boldsymbol{\Theta}_S^{-1}, \quad (31)$$

where $\boldsymbol{\Theta}_C = \boldsymbol{\Theta}_{iq}$ is the diagonal covariance matrix in cepstral domain, $\boldsymbol{\Theta}_S$ the diagonal covariance in the log-spectral domain, and λ a regularization parameter. Note that using the formulation (30) requires an existence of a spectral and a cepstral acoustic model so that each spectral Gaussian is mapped to a specific cepstral Gaussian. The spectral acoustic model can be obtained for example through a forced alignment. The optimization problem in (30) can be solved using the techniques described in Box 1. When minimizing (30), a good starting point is the spectral domain imputation solution given by (25).

Regularization with $\lambda\Theta_S^{-1}$ is necessary when using the cepstral transformation because the matrix $\mathbf{A}^\top\Theta_C^{-1}\mathbf{A}$ is rank-deficient. In [17], an alternative linear transformation was proposed; the ProsPect transformation is a low-order approximation of the cepstral transformation, and like cepstral transformation, it largely uncorrelates the spectral features. The resulting precision matrix \mathbf{P} is, however, full-rank, and no regularization is required. In practice, using ProsPect features is more computationally efficient.

Soft masks

In [23], Gaussian-conditioned imputation was modified to use soft missing data masks. The soft mask $M(t)$ estimated for the t -th frame is represented as a diagonal matrix \mathbf{W} with the mask elements on the diagonal, $\text{diag}(\mathbf{W}) = M(t)$. For soft mask imputation in the cepstral domain, the optimization task in (30) becomes

$$\underset{\xi}{\text{argmin}} \left\{ \frac{1}{2}(\xi - \boldsymbol{\mu}_S)^\top \mathbf{V}(\xi - \boldsymbol{\mu}_S) + \frac{1}{2}(\xi - \boldsymbol{\mu}_S)^\top \mathbf{W}(\xi - \boldsymbol{\mu}_S) \right\} \text{ subject to } \xi \leq \mathbf{X}, \quad (32)$$

where the matrix \mathbf{V} is given as

$$\mathbf{V} = (\mathbf{I} - \mathbf{W})^{\frac{1}{2}} \mathbf{P} (\mathbf{I} - \mathbf{W})^{\frac{1}{2}}, \quad (33)$$

where \mathbf{I} is an identity matrix of the same dimensions as \mathbf{W} . In the formulation (32), the first term ensures that the optimal point gets as close to the Gaussian mean as permitted by the constraint $\xi \leq \mathbf{X}$. The second term that did not exist in the previous formulation (30) ensures that if the mask value is 1 and the features are reliable, the optimal point approaches the observed feature value. Note that the formulation (32) is not limited to the cepstral domain but is equally valid in the ProsPect or log-spectral domain. For log-spectral domain, a closed-form solution of (32) was presented in [23].

0.6 Sparse imputation

0.6.1 Fundamentals

The feature reconstruction methods described in the previous sections are parametric methods that rely on a statistical description of the clean speech characteristics. The front-end based sparse imputation method described in this section, on the other hand, is an exemplar-based feature reconstruction method. Exemplar-based methods model speech using a collection of actual speech samples, *exemplars*. The exemplars typically span several frames which allows the estimation to benefit from temporal correlations. Sparse imputation, first proposed in [24], method works by first finding a small subset of clean speech exemplars that sparsely represent the reliable features of the observed noisy speech. This sparse representation is then used to make an estimate of the unreliable features of the noisy speech.

A sparse representation of clean speech

Let us first consider an utterance that contains only clean speech. We reshape the log-power spectrogram of clean speech, \mathbf{S} , to a single vector s of dimension $L = F \cdot T$ by concatenating

the T subsequent F -dimensional time frames. For now, we assume that T is fixed. We assume that s can be represented exactly or approximated with sufficient accuracy by a linear, non-negative, combination of exemplar spectrograms d_n , where n denotes a specific exemplar ($1 \leq n \leq N$) in the dictionary which contains N available exemplars:

$$s = \sum_{n=1}^N y_n d_n = \mathbf{D}y \quad \text{subject to} \quad y \geq 0 \quad (34)$$

with y an N -dimensional activation vector. The vector y is referred to as a sparse representation of s . The matrix \mathbf{D} denotes the dictionary: $\mathbf{D} = [d_1 \ d_2 \ \dots \ d_N]$, with dimensions $L \times N$ and with $N \gg L$.

If the dictionary is large, the system of linear equations in (34) typically has no unique solution. However, research in the field of Compressive Sensing [25, 26, 27] has shown that if a sparse representation exists, y can be recovered uniquely by enforcing sparsity. Conceptually, sparsity is important because it avoids over-fitting of y and forces the exemplars that are selected to be closer to the underlying, lower-dimensional manifolds on which the various speech classes are located [10]. One possible formulation to obtain a sparse solution is:

$$y = \underset{\tilde{y} \in \mathbb{R}^N}{\operatorname{argmin}} \{ \|\mathbf{D}\tilde{y} - s\|_2 + \lambda \|\tilde{y}\|_1 \} \quad \text{subject to} \quad \tilde{y} \geq 0 \quad (35)$$

with a regularization parameter λ .

Feature reconstruction

If the speech spectrogram contains missing values, we begin by concatenating subsequent time frames of the spectrographic mask \mathbf{M} to form a mask vector m . Using the same approach for the noisy speech spectrogram \mathbf{X} we construct a noisy observation vector x . As for x , its reliable and unreliable elements are denoted m_r and x_u , respectively. We use the reliable elements x_r as an approximation for the corresponding elements of the now unknown s , so problem (35) becomes:

$$y = \underset{\tilde{y} \in \mathbb{R}^N}{\operatorname{argmin}} \{ \|\mathbf{D}_r \tilde{y} - x_r\|_2 + \lambda \|\tilde{y}\|_1 \} \quad \text{subject to} \quad \tilde{y} \geq 0 \quad (36)$$

with \mathbf{D}_r pertaining to the rows of \mathbf{D} for which $m = 1$. The sparse representation y can now be used to estimate the clean observation vector as $\hat{s} = \mathbf{D}y$. In practice, the estimates obtained after solving (36) have some reconstruction error, so it is better to only impute the unreliable elements. Additionally, *bounded* imputation can be approximated by rejecting those elements of which we are sure they have been estimated incorrectly because the estimate exceeds the observed noisy speech:

$$\hat{s} = \begin{cases} \hat{s}_r = x_r \\ \hat{s}_u = \min(\mathbf{D}_u y, x_u) \end{cases} \quad (37)$$

with \mathbf{D}_u and \hat{s}_u pertaining to the rows of \mathbf{D} and \hat{s} for which $m = 0$ and with the *min*-operator taking the element-wise minimum of two values. The denoised spectrogram $\hat{\mathbf{S}}$ is obtained by reshaping \hat{s} into a $F \times T$ matrix.

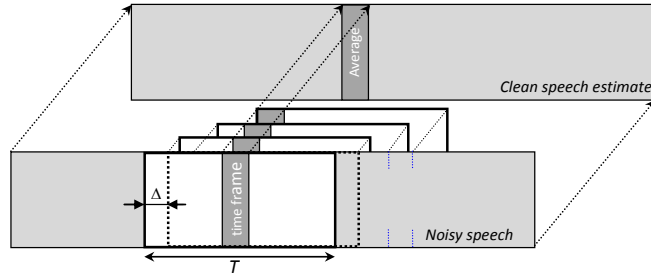


Figure 2: Schematic diagram of the sliding window approach for imputation. The dark shaded time-frame in the noisy utterance is processed in several fixed-length imputation windows, of which we have shown three. Within each window, the given frame takes a different position due to the window shift Δ . The corresponding time-frame in the clean speech estimate is the combination of these individual window-based imputations.

0.6.2 Implementation

Continuous speech

The approach described above is suitable for imputation of noisy speech tokens that can be adequately represented by a fixed number of time frames T [10]. Since arbitrary length utterances clearly do not satisfy this constraint, it is necessary to modify the method. A practical solution was proposed in [28] in the form of a sliding window approach. When using the sliding window approach, the utterance is divided several, overlapping windows by sliding a window of length T through the noisy utterance with shifts of Δ , $1 \leq \Delta \leq T$ frames (cf. Fig. 2). Each window is then imputed separately using sparse imputation as described in Section 0.6.1. Finally, at every time frame, the different clean speech estimates resulting from any overlapping windows are combined.

This recombination can for example be done through averaging or by taking the median value, taking care that only clean speech estimates are used that originate from windows with a non-zero number of reliable elements. If for a certain frame none of the underlying windows contained any reliable features, the sparse imputation method cannot provide a clean speech estimate. If this happens, the clean speech estimate should either not be provided (frame dropping) or it should be calculated based on a different approach such as inserting silence or interpolation.

In [28], it was found that using larger step sizes Δ reduces computational effort but can decrease imputation accuracy. In most subsequent work on sparse imputation, $\Delta = 1$ has been used. The optimal value of the window length T , which translates directly to the length of the exemplars in the dictionary, is database dependent and should be tuned. For AURORA-2, a connected digit database, it was found that $T = 35$ frames of 10ms were optimal [28], while for the large vocabulary SPEECON database, an optimal value $T \approx 20$ frames of 8ms was reported [19].

Creating a dictionary

Sparse imputation models clean speech as a collection of exemplars, the exemplar dictionary. In the case of small, restricted databases such as those available for small vocabulary isolated word recognition, the dictionary can be formed by using all the time-normalized training tokens. When the size of the training database increases, the size of such a dictionary could become impractical. In this scenario, the exemplars should be subsampled from the complete database, for which several options exist. These include clustering, self-organizing maps, and random sampling.

In [10], a single-digit dictionary containing 4000 exemplars was created through random sampling, which yielded an acceptable accuracy at a negligible computational cost. Random sampling ensures a good average coverage of the database, but may not cover under-represented spectra.

When using a sliding window approach, it is probably more important to have shifted variants of exemplars in order to provide shift-invariance. These shifted variants can either be obtained by artificially shifting extracted exemplars, or through sampling with a shifted offset. In [28], a continuous-digit dictionary containing 4000 exemplars was constructed by extracting fixed-size exemplars with a random offset from each utterance in the database. In [19], a dictionary of 8000 exemplars was extracted. In both works, it was reported that while a larger randomly extracted dictionaries may improve performance, the gains are diminishing.

Finding a sparse representation

The computational and algorithmical complexity of sparse imputation is mainly carried by the minimization (36). As minimization using a sparsity constraint has gained considerable interest over the past decade, many off-the-shelf implementations exist. We refer the reader to [29] for an overview and discussion of implementations in various programming languages.

To date, sparse imputation has been used with two different solvers: the basis pursuit interior-point method `l1_ls_nonneg`³ [30], and the greedy `SolveLasso`⁴ solver. For `l1_ls_nonneg` good results were obtained using the default settings, e.g. using the utility function `find_lambda_max_l1_ls_nonneg` to determine the regularization parameter λ and as a stopping criterion a duality gap of 0.01. The `SolveLasso` solver was found to perform better if the algorithm was terminated after 30 iterations. The number of iterations that is optimal depends on the used dictionary and should be empirically tuned.

0.6.3 Advances

Soft masks

In [31], an extension to sparse imputation was proposed that allows the use of a soft missing data mask. When using the soft missing data mask \mathbf{M} , formulation (36) becomes a weighted minimization task,

$$y = \underset{\tilde{y} \in \mathbb{R}^N}{\operatorname{argmin}} \{ \|\mathbf{W}\mathbf{D}\tilde{y} - \mathbf{W}x\|_2 + \lambda \|\tilde{y}\|_1 \} \quad \text{subject to} \quad \tilde{y} \geq 0, \quad (38)$$

³This solver is publicly available from http://www.stanford.edu/~boyd/l1_ls/

⁴This solver is implemented as part of the SparseLab toolbox which is publicly available from <http://www.sparselab.stanford.edu>

where \mathbf{W} is a diagonal matrix the elements of which are determined by the soft missing data mask \mathbf{M} . The weights on the diagonal are given as $\text{diag}(\mathbf{W}) = m$. After obtaining y , imputation is done as

$$\hat{s} = \min(\mathbf{D}y, x). \quad (39)$$

In the new formulation (38), using a binary mask would be equivalent to using \mathbf{W} as a row selector picking only those rows of \mathbf{D} and x that are assumed to contain reliable data. In the case of a soft mask, the weights on the diagonal influence the reconstruction error allowed for each spectrographic element.

0.7 Other feature reconstruction methods

Over the past two decades, numerous other methods for missing feature reconstruction have been proposed. While not all of the methods discussed in this have been applied in ASR tasks, they have been applied to reconstruction spectrograms that contain missing features, and could be used for noise robust ASR.

0.7.1 Parametric approaches

In [11], each frame $S(t)$ was modeled as a non-uniform transformation of the previous frame $S(t-1)$. It was assumed that the linear transformation $\mathbf{A}(t, f)$ applied on the patch $\tilde{S}(t, f)$, centered around $S(t, f)$, was selected from a discrete set of transformations $\{\mathbf{A}_i\}$. The active transformations were modeled as a hidden variable in a generative graphical model whose observed nodes correspond to the patches $\tilde{S}(t, f)$. In the presence of missing values, some of the nodes become hidden. Probabilities of the transformations, $P(\mathbf{A}(t, f) = \mathbf{A}_i)$, and of the hidden nodes, $P(\tilde{S}(t, f))$, were inferred from the observed nodes using a modified form of belief propagation. The transition probabilities between neighboring transformations were determined experimentally. While the method was not evaluated experimentally, visual examples of reconstructed spectra were presented.

In [32], clean speech was modeled as an output of a non-linear state-space model (NSSM). Features $S(t)$ were calculated as a non-linear transformation of hidden source vectors $Z(t)$ which, in turn, were calculated as a non-linear transformation of the previous source vectors $Z(t-1)$. A variational Bayesian approach was used to estimate the transformation parameters from clean speech training data. When using the model for reconstruction, the probability distributions of the source variables, $P(Z(t))$, and of the missing components, $P(S(t, f))$, were inferred from the reliable components using the total derivatives approach proposed in [32]. A small-scale experiment was carried out on clean speech samples with missing values, and performance was evaluated based on the mean square error between the reconstructed and the original clean speech features.

Borgström and Alwan [14] proposed an HMM-based missing value estimation method which can utilize the statistical dependencies between time frames, frequency channels, or both. The feature components $S(t, f)$ were quantized and modeled as a tree-structured set of discrete centroids S_i . Each HMM state corresponds to a centroid, and the state output distributions model the observations probabilities $P(X(t)|S(t, f) \mapsto S_i)$, where $S(t, f) \mapsto S_i$ denotes that the underlying clean speech feature $S(t, f)$ is quantized to S_i . The transition probabilities between centroids were learned from quantized clean speech training data and

the observation probability distributions based on the speech data and a local noise estimate. In reconstruction phase, the probability distribution over the hidden states, $P(S(t, f) \mapsto S_q)$, was inferred using the forward-backward algorithm, and the missing values reconstructed as a weighted sum of the centroids S_i . The HMM-based reconstruction method was evaluated on the AURORA-2 connected digit recognition task.

0.7.2 Non-parametric approaches

In [33], a technique related to sparse imputation was used. Like sparse imputation, the method finds a sparse linear combination of dictionary elements using only the reliable features. Unlike in sparse imputation, the method only works on the current time frame, and has artificial dictionary \mathbf{D} formed by the discrete Haar transform. The optimization problem (36) had the additional constraint that $\mathbf{D}_u y \leq x_u$. With the latter modification, the approximation in Equation (37) becomes unnecessary. In experiments on AURORA-2, performance comparable to the performance of sparse imputation as discussed in Section 0.6 was obtained when estimated mask were used, but oracle mask performance was significantly lower.

In [12], a feature reconstruction method based on non-negative matrix factorization (NMF) was presented. A logarithm-compressed spectrogram \mathbf{S} was represented as a factorization of two matrices, $\mathbf{S} = \mathbf{D}\mathbf{Y}$, where \mathbf{D} describes the spectral envelope templates and \mathbf{Y} describes the power envelopes in time. The approach is related to the sparse imputation method discussed in Section 0.6, with the difference that the spectral dictionary matrix \mathbf{D} is also derived at run-time based on the observed features. The authors described a modification that allows factorization in the presence of missing data, after which reconstruction is done by multiplying the recovered factorizations. Results were reported using SNR and segmental SNR computed on music samples.

The technique proposed in [13] is similar to the NMF-based approach [12] in that spectrograms are described using a latent-variable decomposition. The spectral vectors are, however, expressed in the magnitude rather than log-power domain, which allows the modeling of multiple additive sources. Moreover, for reconstructing features with missing components, the authors use a spectral basis that has been pre-trained using a similar dataset which does not contain missing features. For comparison, the authors used two missing feature reconstruction methods normally applied in different fields: nearest neighbors imputation for computer vision [34] and singular value decomposition (SVD) for imputation of gene expression arrays [35]. The methods were evaluated using visual examples and informal listening tests on music samples.

0.8 Results

The missing feature reconstruction methods described in this chapter have been evaluated in a variety of speech recognition tasks ranging from small vocabulary, isolated word experiments with artificially added noise to large vocabulary continuous speech tasks recorded in realistic environments. However, even when the methods have been evaluated on the same database, the results are difficult to compare because the choice of features, preprocessing, speech recognizer, and missing data mask estimation method all influence the results. To make comparison easier, in this section, the results from various publications are discussed.

In Section 0.8.1, we review results from experiments where imputation methods have been compared against each other. In Section 0.8.2, we discuss the effectiveness of missing feature reconstruction compared to missing feature marginalization and other methods for noise robust speech recognition. In Section 0.8.2, we discuss the influence of mask estimation quality and the effectiveness of using soft masks, and finally, in Section 0.8.3, we discuss results obtained by combining missing feature reconstruction with other techniques such as filtering or multi-condition training.

0.8.1 Imputation methods compared

Raj et al. [6] compared correlation-based, cluster-based, and state-conditioned imputation. Recognition experiments were conducted on the DARPA resource management [36] data which was artificially mixed with white noise and music at a range of SNRs. When oracle masks and spectral features were used, correlation-based and state-conditioned imputation had a comparable performance while cluster-based imputation performed much better. When estimated masks were used, cluster-based and state-conditioned imputation had a comparable performance while correlation-based imputation performed much worse. When the reconstructed features were converted to cepstral domain, the accuracy obtained with cluster-based imputation, and to a lesser extent, with correlation-based imputation, increased substantially. At the time, there was no formulation for class-conditioned imputation in the cepstral domain.

In [9, 17, 37, 38] several modifications of Gaussian-conditioned imputation were compared. The authors carried out experiments on AURORA-2 and AURORA-4 which are databases for noise robust speech recognition research and contain speech artificially corrupted with a variety of noises. AURORA-2 [39] is based on the digit recognition task TIDIGITS [40] and AURORA-4 [41] is a large vocabulary task based on read sentences from the Wall Street Journal (WSJ) [42]. In accordance with the findings in [6], imputation accuracy in the cepstral domain was found superior to the accuracy obtained using the spectral domain. A new linear transformation referred to as ProsPect transformation was proposed in [9] and shown to result in recognition accuracies comparable to the cepstral transformation at a fraction of the computational cost. Moreover, Gaussian-conditioned imputation was modified to allow maximum likelihood channel compensation, which improved accuracy in the presence of a channel mismatch [37].

Faubel et al. [20] compared using cluster-based imputation with minimum mean square error (MMSE) estimates computed using full covariances and bounded MMSE estimates calculated using either diagonal covariances or the full covariance approximation proposed in [20] (cf. Section 0.4.3). Experiments were carried out on the WSJ large vocabulary speech data artificially mixed with noise samples from the NOISEX-92 database. Regardless to whether oracle or estimated masks were used, bounded imputation outperformed unbounded imputation and using full covariance matrices improved the results further still.

In [10], sparse imputation was compared with cluster-based and Gaussian-conditioned imputation. Experiments were carried out on isolated digits from AURORA-2, and reconstructed features were converted to the ProsPect domain prior to recognition. For estimated masks, it was found that sparse imputation performed comparably or somewhat worse than Gaussian-conditioned imputation but much better than cluster-based imputation.

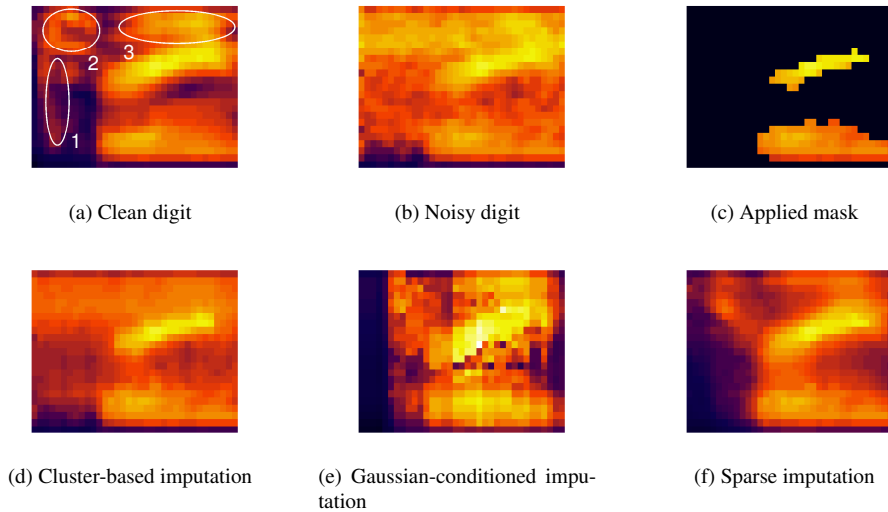


Figure 3: The noisy spectrogram of digit ‘three’ (/θri/) reconstructed using (d) cluster-based imputation, (e) Gaussian-conditioned imputation, and (f) sparse imputation. Comparing the clean speech spectrogram (a) with the noisy observations (b) and the remaining reliable components shown in (c), we see that the imputation method needs to reconstruct 1) the onset, which is a moderate energy pattern seen on the left of the spectrogram, 2) the frication of the /θ/, which is the high energy pattern in the upper left corner, and 3) the formant trace, which is the high energy structure in the upper right corner. The three methods succeed with a varying degree.

For oracle masks, sparse imputation outperformed both Gaussian-conditioned and cluster-based imputation by a large margin at SNRs < 15 dB. The reconstruction results are compared visually in Figure 3. Recognition accuracies obtained with the reconstructed features are reported in Figure 4.

In [19], sparse imputation and cluster-based reconstruction were compared using the Finnish SPEECON database. Experiments on read sentences artificially mixed with babble noise with the reconstructed features converted to cepstra showed that sparse imputation performs much better than cluster-based imputation if an oracle mask was used. When using an estimated mask, sparse imputation performed comparably or slightly better than cluster-based imputation in all but the cleanest conditions. The findings were confirmed in experiments on noisy speech recorded in real-world car and public environments.

0.8.2 Feature reconstruction versus other methods

Traditionally, missing feature reconstruction methods have been compared with the marginalization approach discussed in Chapter ???. In [8, 2], experiments were conducted on artificially noisified TIDIGITS material to compare bounded and unbounded state-conditioned imputation with bounded and unbounded marginalization. The results indicated

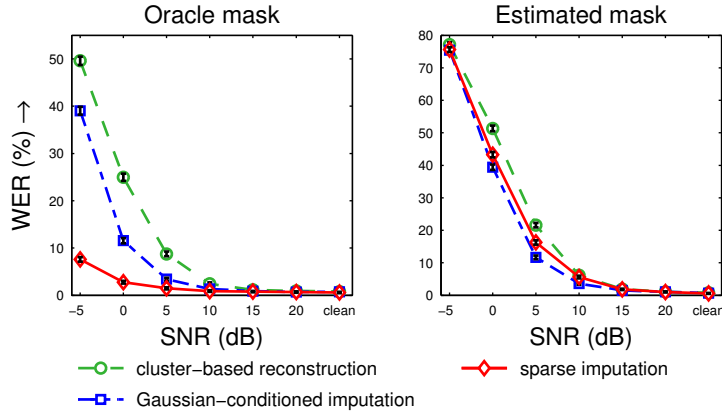


Figure 4: Recognition accuracies obtained on AURORA-2 isolated digits database with cluster-based, Gaussian-conditioned, and sparse imputation. The left panel shows the results obtained using an oracle masks and the right panel shows the results obtained using estimated masks. The horizontal axis describes the SNR at which the clean speech is mixed with the background noise and the vertical axis describes the recognition accuracy averaged over four noise types: subway, car, babble, and exhibition hall noise. The vertical bars around data points indicate the 95 % confidence intervals.

that bounded marginalization works better than bounded imputation, and both work better than their unbounded variants. The experiments did not, however, compare performance on the cepstral features that have been commonly used in later implementations of class-conditioned imputation.

In [6], several feature reconstruction methods were compared with marginalization. It was concluded that marginalization works better than feature reconstruction when used in the spectral domain, but cluster-based reconstruction works better than marginalization when the reconstructed features are transformed to cepstral domain. It was also shown that missing feature techniques perform better than simple spectral subtraction ??.

In [9], cepstral-based Gaussian-conditioned imputation was compared with the ETSI advanced front-end feature extraction (AFE) method [43]. AFE is based on a two-stage Wiener filter approach and is considered a good front-end based feature enhancement tool. Experiments conducted on AURORA-2 showed that the imputation performance is comparable to AFE when an estimated missing data mask is used.

In [44], ProsPect-based Gaussian-conditioned imputation was evaluated in a more challenging task: Flemish SPEECON material that contains noisy speech recorded in real-world environments. In this work, imputation using a clean speech model performed comparably to recognition with a multi-conditioned trained acoustic model. When Gaussian-conditioned imputation was employed with a multi-condition trained acoustic model (cf. 0.8.3), the results were substantially better than those obtained using AFE. A graphical representation of these results can be found in Fig. 5.

Soft estimated masks

While MFT offers an attractive two-stage approach to noise robust ASR, it also makes performance sensitive to mask estimation errors. To show the effect of the imputation algorithm irrespective of possible mask estimation errors, experiments are often conducted using both oracle and estimated masks. While showing the ‘potential’ of the algorithms, it has also made clear that there is a large performance gap between the use of oracle and estimated masks. This is especially true for the sparse imputation method, for which the oracle mask performance is very good, even at low SNRs, but comparable to Gaussian-conditioned imputation when estimated masks are used. In general, it has been found that classifier compensation methods such as conditioned imputation and marginalization are less influenced by mask estimation errors than purely front-end based methods such as cluster-based imputation. [21, 10].

To alleviate the effect of mask estimation errors, it is possible to use a probabilistic rather than binary mask. Making soft decisions was first proposed in [16], where the soft masks were used in missing data marginalization. Experiments on the noisified TIDIGITS database showed that using soft masks substantially increased speech recognition accuracy.

In [21], the effectiveness of using soft masks with cluster-based imputation was investigated. The soft mask cluster-based MMSE estimates were calculated as discussed in Section 0.4.3 and experiments were conducted on a Spanish telephone speech database artificially corrupted with traffic, music, babble and subway noise. Using soft masks resulted in significantly improved recognition performance, particularly for traffic and subway noises, where a 25% relative improvement at 0dB was reported. The soft mask cluster-based imputation performed better than marginalization using a soft mask, although the difference between soft marginalization and imputation methods was observed to decrease at very low SNRs.

In [23], a modification to the Gaussian-conditioned imputation was proposed that allows using soft masks in the spectral or ProsPect domain (c.f. Section 0.5.3). Experiments on AURORA-2 indicated that while using soft rather than binary masks improves the speech recognition performance, the effect is greater when working in spectral domain. Interestingly, the authors also investigated the effectiveness of using a soft version of the oracle mask, and showed that even for an oracle mask, making soft decisions increases the recognition accuracy, although not as much as for estimated masks. This is due to the fact then when the magnitudes of the underlying speech and noise energies are comparable, the approximation in (1) is less accurate.

Finally, in [45] the soft masking approach for sparse imputation, described in Section 0.6.3, was evaluated on AURORA-2. As for the other imputation methods employing soft masks, performance benefited from the probabilistic information in the soft mask. Interestingly, for sparse imputation, the improvement was even larger with oracle masks than with estimated masks: an impressive 8% word error rate was obtained at -5dB when using soft oracle masks.

0.8.3 *Combination with other methods*

A number of authors have proposed to combine missing feature reconstruction with other feature enhancement methods. Combining missing data techniques with spectral subtraction was first proposed in [46], where marginalization combined with spectral subtraction was

used in speaker verification task. In [6], spectral subtraction was used on the reliable features of speech artificially corrupted with white noise, and correlation-based, cluster-based, and state-conditioned imputation were all shown to benefit from spectral subtraction. Gaussian-conditioned imputation was applied on spectral subtracted speech in [44] with similar results. In that work, the authors concluded that the performance improves because the imputation bounds became more accurate.

In [20, 47], cluster-based imputation was combined with a particle filtering technique that had previously been employed in feature enhancement [48, 49]. Particle filtering was used for calculating estimates of the underlying clean speech and noise. The estimates were used twice: first, the clean speech and noise estimates were used to construct a missing data mask, and then, cluster-based imputation was applied on the clean speech estimate, guided by this missing data mask. Investigations on artificially noisified WSJ large vocabulary speech data indicated that both particle filtering and missing feature reconstruction contributed in the improved speech recognition performance.

Finally, in [44] using Gaussian-conditioned imputation in combination with multi-condition trained acoustic models was proposed. In theory, the combination is incorrect since the assumption that reliable features remain effectively uncorrupted means the reliable features should be reconstructed and recognized using a clean speech model. Figure 5 shows, however, that in practice, the combination leads to substantial improvements in recognition accuracy. The experiments were conducted on noisy speech from the SPEECON database which has been recorded in realistic conditions. It was concluded that using a multi-condition trained model describes a wider variance of speech phenomena and thus compensates not only for additional effects such as reverberation but also for mask estimation errors. After all, if a unreliable feature is erroneously labeled reliable, the noisy speech model has a better chance of recovering from the mask estimation error.

0.9 Discussion and conclusion

We have discussed several methods for feature reconstruction as an approach to improve noise robustness in ASR under the missing data paradigm. Four well-known methods, namely correlation-based imputation, cluster-based imputation, class-conditioned imputation, and sparse imputation, were discussed in detail along with some significant advances that have been proposed to improve the basic approach. The performance of the methods was analyzed based on results gathered from various studies. Additionally, a number of recently developed methods that have not been extensively evaluated in noisy speech recognition task were described in Section 0.7.

The results discussed in Section 0.8.1 suggest that the most effective feature reconstruction methods are sparse imputation and Gaussian-conditioned imputation. Sparse imputation typically results in the best speech recognition performance when oracle masks are used whereas Gaussian-conditioned imputation results in the best performance when estimated masks are used. However, while effective, Gaussian-conditioned imputation is not a front-end based method, and requires classifier modification, especially when using the advances outlined in Section 0.5.3.

If a front-end reconstruction method is preferred, sparse imputation or cluster-based imputation are recommended, for correlation-based imputation has performed worse than cluster-based imputation in all recorded experiments on realistic data. Cluster-based

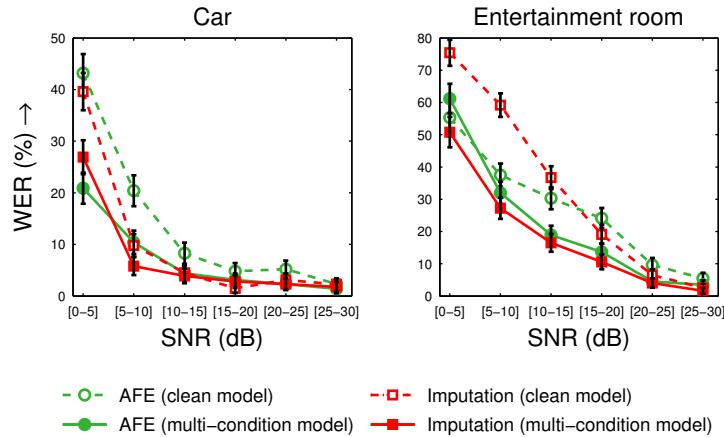


Figure 5: Recognition accuracies obtained with Gaussian-conditioned imputation and the ETSI AFE feature enhancement method when either a clean speech or multi-condition trained acoustic model was used. Evaluated on noisy speech recorded in a car environment, the choice of acoustic model does not affect the order of the methods, but imputation outperforms AFE in both cases. In the more challenging entertainment room environment, imputation outperforms AFE only if the multi-condition model is employed. Using the multi-condition model generally improves the results in both environments.

imputation, on the other hand, was shown to be as effective as sparse imputation when the noise level is moderate and estimated masks are used. While the result may depend on the recognition task, cluster-based imputation seems a fair alternative for low-noise conditions and it is easy to implement.

The comparisons between missing feature reconstruction and noise robustness methods not based on MDT, reviewed in Section 0.8.2, indicated that feature reconstruction can result in a performance as good as or better than multi-condition training, spectral subtraction, or feature enhancement with the ETSI AFE frontend. When compared with marginalization, the feature reconstruction methods appeared to work better when the reconstructed features were transformed to cepstral domain prior to recognition, which is not possible with standard marginalization approaches. Marginalization has been extended to cepstral domain, but the cepstral domain marginalization approach has not been compared with feature reconstruction approaches [50].

It is also unfortunate that no feature reconstruction method has been directly compared with popular model-based noise robustness methods such as parallel model combination (PMC) or vector Taylor series (VTS) approximation [51, 52, 53]. Comparing recognition accuracies reported on databases such as AURORA-2 and AURORA-4 that have been used for evaluating all methods, seem to indicate that feature reconstruction could only outperform model-based compensation methods when oracle masks are used. When estimated masks are used, Gaussian-conditioned imputation, which is the most effective reconstruction method in this setting, performs comparable to or somewhat worse than model-based noise compensation.

The large difference between oracle mask and estimated mask performance of sparse imputation exemplifies how the performance of feature reconstruction methods is largely determined by the quality of the missing data mask. Depending on the data, estimating the missing data mask with a sufficient accuracy can be extremely difficult. Although the results discussed in Section 0.8.2 showed that performance improves when soft masks are used, using soft masks does not bridge the gap between oracle and estimated mask performance. Moreover, experiments on more challenging data such as noisy speech recorded in realistic environments indicated that missing feature reconstruction may be more difficult if the additivity assumptions of noise and speech are violated, which happens, for example, in the presence of reverberation [44].

The results discussed in Section 0.8 suggest that, in general, the performance of feature reconstruction methods improves as more information is used for missing value estimation. That is, bounded imputation works better than unbounded imputation as reported in [2] and using the uncertainties encoded in soft masks improves the results over the binary mask results as discussed in Section 0.8.2. And, especially in noisy conditions, using a time context that spans more than a single frame has been shown to be important for sparse imputation [19] and HMM-based reconstruction [14].

Finally, the results discussed in Section 0.8.3 show that combining feature reconstruction with other noise robustness techniques leads to improved speech recognition performance. This, together with the fact that mask estimation allows integrating additional sources of knowledge (e.g. harmonicity) in the noise compensation process, means there is still potential for improving feature reconstruction performance. Interesting results could arise from combining feature reconstruction with more advanced noise robustness techniques, such as the ETSI AFE front-end or VTS approaches.

References

- [1] M. Cooke, P. Green, and M. Crawford, "Handling missing data in speech recognition," in *Proc. International Conference on Speech and Language Processing*, 1994, pp. 1555–1558.
- [2] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, June 2001.
- [3] J. Barker, M. Cooke, and P. Green, "Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *Proc. EUROSPEECH*, Aalborg, Denmark, September 3–7 2001, pp. 213–216.
- [4] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition," *Proc. International Conference on Audio, Speech and Signal Processing*, vol. 2, pp. 863–866, 1997.
- [5] B. Raj, "Reconstruction of incomplete spectrograms for robust speech recognition," Ph.D. dissertation, Carnegie Mellon University, 2000.
- [6] B. Raj, M. Seltzer, and R. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275–296, September 2004.
- [7] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
- [8] L. Josifovski, M. Cooke, P. Green, and A. Vizinho, "State based imputation of missing data for robust speech recognition and speech enhancement," in *Proc. EUROSPEECH*, Budapest, Hungary, September 5–9 1999, pp. 2837–2840.
- [9] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. International Conference on Audio, Speech and Signal Processing*, vol. 1, 2004, pp. 213–216.

- [10] J. F. Gemmeke, H. Van hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 272–287, 2010.
- [11] M. J. Reyes-Gomez, N. Jovic, and D. P. Ellis, "Towards single-channel unsupervised source separation of speech mixtures: The layered harmonics/formants separation/tracking model," in *ISCA tutorial and research workshop on statistical and perceptual audition (SAPA)*, 2004.
- [12] J. L. Roux and A. de Cheveigne, "Computational auditory induction by missing-data non-negative matrix factorization," in *ISCA tutorial and research workshop on statistical and perceptual audition (SAPA)*, 2008.
- [13] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for spectral audio signals," in *IEEE Intl. workshop on machine learning for signal processing*, 2009.
- [14] B. Borgström and A. Alwan, "HMM-based reconstruction of unreliable spectrographic data for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1612–1623, 2010.
- [15] A. Nadas, D. Nahamoo, and M. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, October 1989.
- [16] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. International Conference on Speech and Language Processing*, Beijing, China, 2000, pp. 373–376.
- [17] H. Van hamme, "Prospect features and their application to missing data techniques for robust speech recognition," in *Proc. INTERSPEECH*, 2004, pp. 101–104.
- [18] F. Sha, L. K. Saul, and D. D. Lee, "Multiplicative updates for nonnegative quadratic programming in support vector machines," in *Proc. Neural Information Processing Systems*. MIT Press, 2002, pp. 1041–1048.
- [19] J. F. Gemmeke, B. Cranen, and U. Remes, "Sparse imputation for large vocabulary noise robust ASR," *Computer Speech & Language*, vol. 25, no. 2, pp. 462–479, 2011.
- [20] F. Faubel, J. McDonough, and D. Klakow, "Bounded conditional mean imputation with Gaussian mixture models : a reconstruction approach to partly occluded features," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2009, pp. 3869–3872.
- [21] B. Raj and R. Singh, "Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 65–70.
- [22] H. Van hamme, "Robust speech recognition using missing feature theory in the cepstral or LDA domain," in *Proc. of European Conference on Speech Communication and Technology*, 2003, pp. 3089–3092.
- [23] M. Van Segbroeck and H. Van hamme, "Robust speech recognition using missing data techniques in the PROSPECT domain and fuzzy masks," in *Proc. International Conference on Audio, Speech and Signal Processing*, March 30–April 4 2008, pp. 4393–4396.
- [24] J. F. Gemmeke and B. Cranen, "Using sparse representations for missing data imputation in noise robust speech recognition," in *Proc. EUSIPCO*, Lausanne, Switzerland, August 25–29 2008.
- [25] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.
- [26] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, June 2006.
- [27] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications On Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, March 2006.
- [28] J. F. Gemmeke and B. Cranen, "Missing data imputation using compressive sensing techniques for connected digit recognition," in *Proc. DSP*, Santorini, Greece, July 5–7 2009, pp. 1–8.
- [29] I. Carron, "Compressive Sensing: The Big Picture [Internet]," <http://sites.google.com/site/igorcarron2/cs>, 2009.
- [30] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l_1 -regularized least squares," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, December 2007.
- [31] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2010.
- [32] T. Raiko, M. Tormio, A. Honkela, and J. Karhunen., "State inference in variational bayesian nonlinear state-space models," in *Proc. International Conference on Independent Component Analysis and Blind Source Separation*, 2006, pp. 222–229.

- [33] B. Borgström and A. Alwan, "Utilizing compressibility in reconstructing spectrographic data with applications to noise robust asr," *Signal Processing Letters*, vol. 16, no. 5, pp. 398–401, 2009.
- [34] M. E. Brand, "Incremental singular value decomposition of uncertain data with missing values," in *Proc. European Conference on Computer Vision (ECCV)*, vol. 2350, 2002, pp. 707–720.
- [35] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, "Imputing missing data for gene expression arrays," Stanford Statistics Department, Tech. Rep., 1999.
- [36] P. Price, W. Fisher, J. Bernstein, and D. Pallet, "State inference in variational bayesian nonlinear state-space models," in *Proc. IEEE Conf. on Acoustics Speech and Signal Processing*, 1998, pp. 651–654.
- [37] M. Van Segbroeck and H. Van hamme, "Handling convolutional noise in missing data automatic speech recognition," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2006, pp. 2562–2565.
- [38] M. Van Segbroeck and H. Van hamme, "Advances in missing feature techniques for robust large vocabulary continuous speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 123–137, 2011.
- [39] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA Tutorial and Research Workshop ASR2000*, Paris, France, September 18–20 2000, pp. 181–188.
- [40] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. International Conference on Audio, Speech and Signal Processing*, 1984, pp. 328–331.
- [41] N. Parihar and J. Picone, "An analysis of the aurora large vocabulary evaluation," in *Proc. of Eurospeech*, 2003, pp. 337–340.
- [42] D. Paul and J. Baker, "The design of wall street journal-based CSR corpus," in *Proc. Int. Conf. Spoken Language Systems*, 1992, pp. 899–902.
- [43] "ETSI standard doc.: Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; ES 202 050 V1.1.5," 2007.
- [44] J. F. Gemmeke, M. Van Segbroeck, Y. Wang, B. Cranen, and H. Van hamme, "Automatic speech recognition using missing data techniques: Handling of real-world data," in *To appear in: Robust Speech Recognition of Uncertain or Missing Data*, R. Haeb-Umbach and D. Kolossa, Eds. Springer.
- [45] J. F. Gemmeke and B. Cranen, "Sparse imputation for noise robust speech recognition using soft masks," in *Proc. International Conference on Audio, Speech and Signal Processing*, Taipei, Taiwan, April 19–24 2009, pp. 4645–4648.
- [46] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proc. International Conference on Audio, Speech and Signal Processing*, 1998.
- [47] F. Faubel, H. Raja, J. McDonough, and D. Klakow, "Particle filter based soft-mask estimation for missing feature reconstruction," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2008.
- [48] B. Raj, R. Singh, and R. Stern, "On tracking noise with linear dynamical system models," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2004.
- [49] F. Faubel and M. Wölfel, "Overcoming the vector taylor series approximation in speech feature enhancement - a particle filter approach," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2007.
- [50] J. Häkkinen and H. Haverinen, "On the use of missing feature theory with cepstral features," in *Proc. CRAC Workshop, Aalborg, Denmark*, 2001.
- [51] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, 1996.
- [52] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector taylor series for noise speech recognition," in *Proc. the International Conference on Spoken Language Processing*, Beijing, China, 2000.
- [53] R. C. van Dalen, F. Flego, and M. J. F. Gales, "Transforming features to compensate speech recogniser models for noise," in *Proc. INTERSPEECH*, Brighton, UK, September 6–10 2009, pp. 2499–2502.