Compressive Sensing for Missing Data Imputation in Noise Robust Speech Recognition

Jort Florent Gemmeke*, *Student-Member, IEEE,* Hugo Van hamme, *Member, IEEE,*
Bert Cranen, Lou Boves

Abstract—An effective way to increase the noise robustness of automatic speech recognition is to label noisy speech features as either reliable or unreliable (missing), and to replace (impute) the missing ones by clean speech estimates. Conventional imputation techniques employ parametric models and impute the missing features on a frame-by-frame basis. At low SNR's these techniques fail, because too many time frames may contain few, if any, reliable features.

In this paper we introduce a novel non-parametric, exemplar-based method for reconstructing clean speech from noisy observations, based on techniques from the field of Compressive Sensing. The method, dubbed *sparse imputation*, can impute missing features using larger time windows such as entire words. Using an overcomplete dictionary of clean speech exemplars, the method finds the sparsest combination of exemplars that jointly approximate the reliable features of a noisy utterance. That linear combination of clean speech exemplars is used to replace the missing features.

Recognition experiments on noisy isolated digits show that sparse imputation outperforms conventional imputation techniques at SNR = −5 dB when using an ideal 'oracle' mask. With error-prone estimated masks sparse imputation performs slightly worse than the best conventional technique.

Index Terms—Compressive sensing, missing data techniques, noise robustness, automatic speech recognition.

I. INTRODUCTION

REMOVING a foreground object that partially occludes the image of interest is a well-known image processing task (cf. Fig. 1). Occlusion due to the presence of objects between the camera and the object(s) of interest is a pervasive problem in image recognition. Recognition performance can be improved by discarding the features that are missing due to the occlusion, or by imputing the missing features on the basis of what is still visible [1], [2]. Speech recognition in the presence of competing audio signals can also be formulated as a missing data problem, similar to the treatment of partially occluded images. Audio signals can be represented as two-dimensional grey-scale (or color) pictures, where one axis represents time, the other represents frequency and the grey value (or color) represents the acoustic energy at a specific instant in time in a specific frequency band (cf. Fig. 2a). If the noise power in a certain time-frequency area is larger than the power of the speech, it can be said that the noise occludes or masks the speech. In Automatic Speech Recognition (ASR) Missing Data Techniques (MDTs) [3]–[5] do indeed provide a powerful way to mitigate the impact of both stationary and non-stationary noise for a wide range of Signal-to-Noise ratios (SNR).

Obviously, MDT hinges on the assumption that it is possible to estimate −prior to decoding− which spectro-temporal elements represent speech and which represent background noise that 'occludes' the speech. These estimates, referred to as a *spectrographic mask*, can then be used to instruct the decoder to ignore these elements (known as *marginalization*), or to replace the occluded elements by clean speech estimates prior to or during decoding. The latter case is an example of *missing data imputation* [6], [7]. In this paper we will only investigate imputation techniques.

While missing data imputation appears to be very effective in noise robust ASR at moderate SNR levels ≥ 10 dB, the performance of conventional techniques drops substantially at SNR levels ≤ 0 dB, even when using an 'ideal' spectrographic mask (cf. Fig. 6). This drop is due to several interrelated problems. First, the proportion of data that is missing is substantial: at SNR = −5 dB over 80% of the data needs to be imputed (cf. Fig 3). Second, contrary to the typical case in image recognition, occlusions are not confined to compact regions of the spectro-temporal picture (cf. Fig. 2c). While a random distribution of occlusions might seem conducive to estimating the features of the occluded parts, in actual practice it gives rise to the third problem: It becomes difficult to know which parts of the picture represent speech and which represent noise. The difficulty of telling speech from noise is only aggravated by the fact that (different from most image recognition tasks) even in clean speech there are no sharp boundaries between speech and 'silence'. Finally, the energy in a spectro-temporal cell is a random variable in its own right. A speaker cannot produce the exact same signal twice when repeating a word or an utterance. Moreover, small changes in the position of the microphone relative to the lips and the properties of a specific microphone and transmission channel may result in a large change of acoustic energy.

From the articulation processes that produce speech signals it can be inferred that values of adjacent time-frequency cells are strongly correlated along both axes. Yet, conventional

J.F. Gemmeke (e-mail: J.Gemmeke@let.ru.nl), B. Cranen (email: B.Cranen@let.ru.nl) and L. Boves (email: L.Boves@let.ru.nl) are with the Centre for Language and Speech Technology, Radboud University Nijmegen, NL-6500 HD Nijmegen, The Netherlands

H. Van hamme is with the department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, B-3001 Heverlee, Belgium, (e-mail: hugo.vanhamme@esat.kuleuven.be).

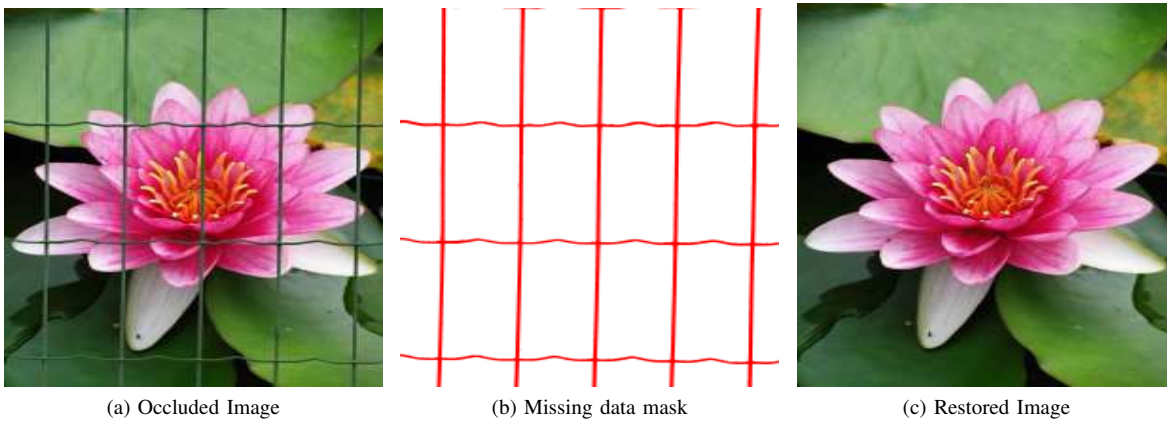| (a) Occluded Image | (b) Missing data mask | (c) Restored Image |

Fig. 1. A typical task in image processing, 'inpainting', is removing a foreground object from an occluded image (Fig. 1a) using a manually selected missing data mask (Fig. 1b), yielding the unoccluded object displayed in Fig. 1c.

imputation techniques for ASR employ parametric models for reconstructing the spectral envelope on a frame-by-frame basis (i.e., for individual time slices). *Parametric* models are used because until recently non-parametric methods for reconstructing spectral envelopes from a possibly small number of 'clean' observations were not available. Imputation is limited to one axis because the number of parameters of models that cover a sufficiently wide window in two dimensions quickly becomes unwieldy [7]. The preference for the frequency axis over the time axis is because in general the spectral envelope is smoother than the time envelope. Yet, limiting the imputation to the spectral envelope of a single time frame makes this approach especially vulnerable in frames that contain few spectral regions where speech energy is higher than the energy of the competing sounds. Here, help from expectations based on the temporal envelope could come in handy. Thus, it would seem unlikely that frame-based parametric techniques for reconstructing clean speech spectra from noisy speech observations can solve the recognition problems at SNR levels ≤ 0 dB.

In this paper we introduce a non-parametric, exemplar-based, method for reconstructing clean speech from noisy observations, based on a *Compressive Sensing* approach [8], [9]. The approach, dubbed *sparse imputation*, can impute missing features using time windows that comprise multiple frames. Conceptually, the use of exemplar-based imputation can be justified with a metaphor: if we observe a few mountain tops above a blanket of low clouds, and we have cloud-free 3-D representations of all mountainous areas on the planet, we can reconstruct the invisible terrain very accurately by finding the representations that match best with the observations. Due to the intrinsic variability in speech *exact* reconstruction of a speech spectrum from a small number of observations may be impossible, but because of the fact that speech signals are observations of a random process to begin with, this is probably not necessary either.

The theory of Compressive Sensing (CS) asserts that if a signal (such as a picture) can be expressed as a sparse linear combination of vectors, it can be recovered using a very limited number of measurements. In [10] it was suggested

that CS techniques can be used for missing data imputation. They illustrated their approach by recovering missing pixels in images that were sparsely represented in an inverse discrete cosine transformation (IDCT) basis. The technique works by treating the non-missing pixels as measurements of an unknown sparse representation. After finding the sparse representation, the complete picture can be recovered by projecting the sparse representation in the IDCT basis. In [11] it was suggested that a picture might be very sparsely represented in an overcomplete dictionary of *examples*, by expressing that picture as a linear combination of a small number of example images.

In this paper we investigate whether a combination of the approaches proposed in [10] and [11] can be applied to noisy speech. Thus, the goal of the paper is to explore whether *sparse imputation* can solve the missing data imputation problems for noise robust ASR that were sketched above. To that end we compare recognition accuracies obtained using sparse imputation with the results obtained with state-of-the-art conventional imputation techniques. As a first step towards more general ASR tasks we test our approach with material from the well-known AURORA-2 digit recognition task [12]. While doing so, we address two issues in particular. First, since the minimum proportion of spectro-temporal features that is required for reconstructing clean speech spectra is not known, we develop a theoretical estimate for this proportion and put it to an experimental test. Second, to investigate the influence of mask estimation errors, we compare two types of masks: 1) The 'oracle' mask[1] and the *harmonicity mask* that derives reliability estimates from a harmonic decomposition [13].

The rest of the paper is organized as follows. In Section II we introduce Missing Data Techniques for ASR and the two types of missing data masks that we will compare. In Section III we describe the sparse imputation framework. In addition, we propose a theoretical estimate for the minimum number of spectro-temporal features that are needed for successful reconstruction of noise-free representations. In Section IV we

---

[1]Oracle masks are masks in which reliability decisions are based on exact knowledge about the extent to which each time-frequency element is dominated by either noise or speech.
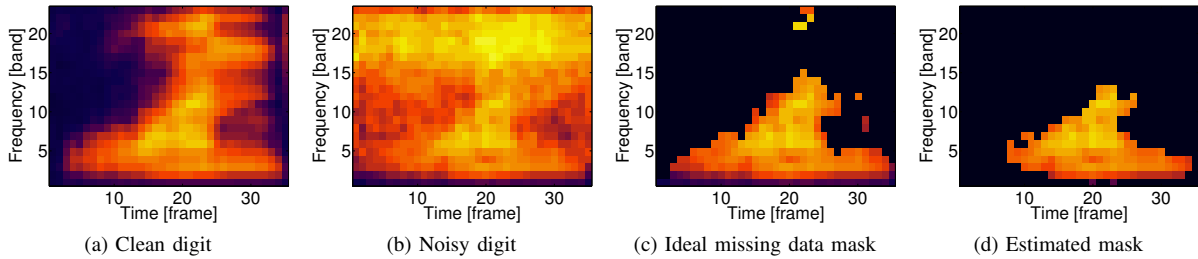
(a) Clean digit   (b) Noisy digit   (c) Ideal missing data mask   (d) Estimated mask

Fig. 2. Fig. 2a shows the spectro-temporal representation of the digit 'one'. In Fig. 2b the clean speech is artificially corrupted by suburban train noise at SNR = −5 dB. The horizontal axis represents time, the vertical axis represents frequency and the intensity represents the acoustic energy. As can be observed in Fig. 2c, a substantial part of the data needs to be imputed even when using an ideal missing data mask which is calculated using knowledge of the corrupting noise. Comparison with the realistic estimated mask in Fig. 2d shows that the mask estimation is not error-free. In this case this results in even more missing data that must be imputed.

briefly describe the two conventional imputation techniques against which the novel sparse imputation technique will be compared. In Section V we explain the design of the experiments and the results are presented in Section VI. We discuss the results in Section VII and suggestions for future research in Section VIII; we present our conclusions in Section IX.

II. MISSING DATA TECHNIQUES IN ASR

A. Motivation

In this Section we give a very brief introduction to the use of MDT for noise robust ASR [14], [15]. In ASR, speech is represented as a spectro-temporal distribution of acoustic power, a *spectrogram*. In noise-free conditions, the value of each time-frequency cell in the spectrogram, a two-dimensional matrix, is determined only by the speech signal. In noisy conditions, the power in each cell represents a combination of speech and background noise.

Assuming noise is additive, the power spectrogram of noisy speech, denoted by Y, can be approximately described as the sum of the individual power spectrograms of clean speech S and noise N, i.e., $Y = S + N$. ASR systems mimic human hearing by employing logarithmic compression resulting in log-spectral energy features. The logarithmic compression of a sum can be approximated by a compression of the largest of the two terms [16]. For noisy speech features in which the speech energy dominates we can write:

$$
\begin{aligned}
\log[S(k,t)+N(k,t)] &= \log[S(k,t)(1+\tfrac{N(k,t)}{S(k,t)})] \\
&\approx \log[S(k,t)]
\end{aligned} \quad (1)
$$

with the spectrograms S, N and Y represented as $K \times T$ dimensional matrices (with K the number of frequency bands and T the number of time frames) indexed by frequency band k $(1 \leq k \leq K)$ and time frame t $(1 \leq t \leq T)$.

From (1) we can infer that noisy speech features in which the speech energy dominates remain approximately uncorrupted and can be used directly as estimates of the clean speech features.

B. Missing data masks

Elements of Y that predominantly contain speech or noise energy are distinguished by introducing a spectrographic mask

M. The elements of a mask M are either 1, meaning that the corresponding element of Y is dominated by speech ('reliable') or 0, meaning that it is dominated by noise ('unreliable' c.q. 'missing'). Thus, we write:

$$
M(k,t) = \begin{cases} 1 \stackrel{def}{=} \text{reliable} & \text{if } \frac{S(k,t)}{N(k,t)} > \theta \\ 0 \stackrel{def}{=} \text{unreliable} & \text{otherwise} \end{cases} \quad (2)
$$

with constant threshold θ. Smaller values of θ will result in more elements considered as reliable in the mask, but the proportion of errors implied in the assumption that $S(k,t) = Y(k,t)$ will be larger, while larger values of θ lead to a safer model, but fewer reliable elements to impute the missing data from.

C. Estimating missing data masks

In experiments with artificially added noise, the *oracle masks* can be computed directly by means of (2) using knowledge of the corrupting noise and the clean speech signal. The oracle mask is useful to assess the potential of missing data imputation techniques and to compare the performances of different techniques in ideal conditions.

In realistic situations, however, the masks must be estimated from the noisy speech. Many different estimation techniques have been proposed, such as SNR based estimators [17], mask estimation by means of Bayesian classifiers [18], [19], methods that focus on speech characteristics, e.g. harmonicity based SNR estimation [13], and mask estimation exploiting binaural cues [20] or correlogram structure [21] (cf. [22] and the references therein for a more complete overview of mask estimation techniques). In the experiments presented in this paper we used the oracle mask and the estimated harmonicity mask [13].

Fig. 3 shows the proportion of missing data in the AURORA-2 database for several SNR values, both for the oracle and the estimated harmonicity mask. The most interesting observations that can be made from that figure are (1) that the harmonicity mask is more biased towards considering spectral values unreliable than the oracle mask, (2) that the proportion of unreliable values varies widely for every SNR value, (3) that the harmonicity mask considers a substantial proportion of the values in clean speech as unreliable, and (4) that even for the oracle mask more than 80% of the data are unreliable at the SNR value -5 dB.
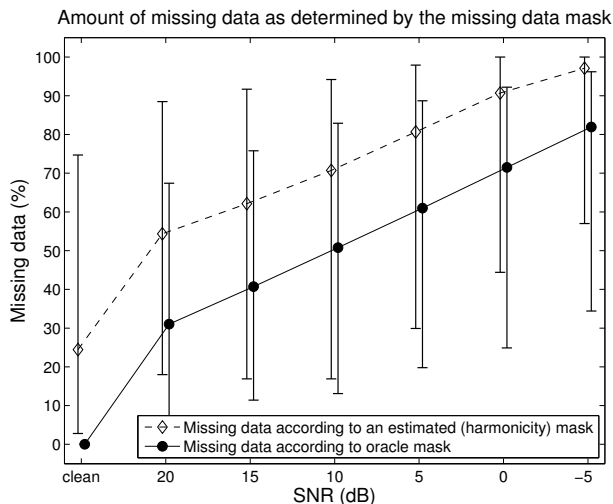
Fig. 3. The percentage of missing data as a function of SNR for all digits in the test database of AURORA-2. Results are shown for the oracle missing data mask, which is calculated from exact knowledge of the corrupting noise, as well as for an estimated mask, the harmonicity mask described in Section V-C. The vertical bars around the data points show the 1st and 99th percentile.

D. Use of MDT in ASR

Techniques for speech recognition in the presence of missing data can be divided in two categories: marginalization and imputation. In the marginalization approach [4], [23] acoustic likelihoods are calculated by integrating over the range of possible values of the missing features and recognition is carried out primarily based on the reliable features. In the imputation approach [6], [7] the missing features are replaced by clean speech estimates, after which recognition can proceed without modification of the recognition system. In *conditional imputation* the clean speech estimates are made dependent on the underlying statistics, such as the hypothesized state.

The advantage of the imputation approach is that the reconstructed clean speech features can be converted to cepstral features, which improves recognition accuracy at high SNR's. Marginalization, on the other hand, has been shown to be more robust against data scarcity at low SNRs than traditional imputation methods [4]. In this paper we will only investigate imputation techniques.

E. Bounded MDT

Both marginalization and imputation approaches are called *unbounded* if there are no restrictions on the range of possible values the unreliable features can take. In this work we consider only additive noise. This implies that the observed acoustic power of noise corrupted speech can be considered as an upper bound for a clean speech estimate:

$$\hat{\boldsymbol{S}} = \begin{cases} \hat{\boldsymbol{S}}(k,t) = \boldsymbol{Y}(k,t) & \text{if } \boldsymbol{M}(k,t) = 1 \\ \hat{\boldsymbol{S}}(k,t) \leq \boldsymbol{Y}(k,t) & \text{if } \boldsymbol{M}(k,t) = 0 \end{cases} \quad (3)$$

In reconstructing the clean speech estimate $\hat{\boldsymbol{S}}$ the upper bound given by (3) should not be exceeded.

III. SPARSE IMPUTATION

A key concept in Compressive Sensing is that many real-life signals have a sparse representation given an appropriate change of basis. In Section III-A we will show how speech signals corresponding to spoken digits can be sparsely represented in a dictionary of example speech tokens and how such a sparse representation can be recovered from observed spectrographic elements. In Section III-B we show how the sparse representation can be recovered from incomplete spectrograms and how the missing data can be reconstructed. In Section III-C we discuss the difficulties associated with determining how much reliable data must be available to reconstruct the spectrogram of a spoken digit in the presence of competing acoustic signals.

A. Sparse representation of speech

We express the $K \times T$ spectrogram of clean speech $\boldsymbol{S}$ as a single vector $\boldsymbol{s}$ of dimension $D = K \cdot T$ by concatenating T subsequent time frames. To keep the correspondence with research in image processing, we assume that T can be fixed. This can be achieved, for example, by time-normalizing all utterances [24].

Inspired by a similar approach in the field of face recognition [11], we assume that $\boldsymbol{s}$ can be represented exactly (or at least approximated with sufficient accuracy) by a linear combination of exemplar spectrograms $\boldsymbol{a}_n$, where n denotes a specific exemplar $(1 \leq n \leq N)$ in the set of N available exemplars:

$$\boldsymbol{s} = \sum_{n=1}^{N} x_n \boldsymbol{a_n} = \boldsymbol{A}\boldsymbol{x} \quad (4)$$

with $\boldsymbol{x}$ an N-dimensional weight vector,[2] and the overcomplete dictionary $\boldsymbol{A} = (\boldsymbol{a_1} \quad \boldsymbol{a_2} \ldots \boldsymbol{a_N})$ a matrix of size $D \times N$ with $N \gg D$. In fact, since the dictionary is overcomplete, any vector can be represented as a linear combination of vectors from the dictionary.

Although it may not be obvious at first that an arbitrary log-power spectrogram can be represented as a sparse linear combination of similar spectrograms, the experimental data below indicates that this is a reasonable assumption. The reason for this is that spectrograms of different realizations of the same word have approximately the same patterns of energy concentration. The differences between multiple exemplar spectrograms of the same word manifest themselves mainly as relatively small variations in the shape and position of the high-energy regions in the time-frequency plane. As a consequence, a linear combination of exemplar spectrograms that represent the same word, will result in a new spectrogram that looks very similar to a possible realization of that word but with slightly different boundaries of the high-energy regions.

Although the system of linear equations in (4) has no unique solution, research in the field of Compressive Sensing [8], [9]

---

[2] We do not require that $\boldsymbol{x}$ is non-negative. In practice, however, we hardly observe any negative values.

has shown that if $\boldsymbol{x}$ is sufficiently *sparse*, $\boldsymbol{x}$ can be *uniquely* determined by solving:

$$\boldsymbol{x} = \operatorname*{argmin}_{\tilde{\boldsymbol{x}} \in \mathbb{R}^N} \{ \|\tilde{\boldsymbol{x}}\|_0 \} \text{ subject to } \boldsymbol{s} = \boldsymbol{A}\tilde{\boldsymbol{x}} \quad (5)$$

with $\|.\|_0$ the l^0 zero norm (i.e., the number of nonzero elements).

The combinatorial problem (5) is NP-hard [25] and therefore unfeasible for practical applications. It has been shown in [26] however, that with weak conditions on $\boldsymbol{A}$ the solution of the l^0 zero norm minimization is equal to the solution of an l^1 norm minimization:

$$\boldsymbol{x} = \operatorname*{argmin}_{\tilde{\boldsymbol{x}} \in \mathbb{R}^N} \{ \|\tilde{\boldsymbol{x}}\|_1 \} \text{ subject to } \boldsymbol{s} = \boldsymbol{A}\tilde{\boldsymbol{x}} \quad (6)$$

This convex minimization problem can be cast as a least squares problem with an l^1 penalty, also referred to as the LASSO [27]:

$$\boldsymbol{x} = \operatorname*{argmin}_{\tilde{\boldsymbol{x}} \in \mathbb{R}^N} \{ \|\boldsymbol{A}\tilde{\boldsymbol{x}} - \boldsymbol{s}\|_2 + \lambda \|\tilde{\boldsymbol{x}}\|_1 \} \quad (7)$$

with a regularization parameter λ. Public domain software packages exist to solve problem (7) efficiently.

We can use this approach to obtain a sparse representation $\boldsymbol{x}$ of the clean speech vector $\boldsymbol{s}$ by treating the speech features as measurements of the unknown sparse signal $\boldsymbol{x}$.

B. Imputation

By concatenating subsequent time frames of the spectrographic mask $\boldsymbol{M}$, similarly as we did for the clean speech spectrogram $\boldsymbol{S}$, we construct a mask vector $\boldsymbol{m}$. Using the same approach for the noisy speech spectrogram $\boldsymbol{Y}$ we construct a noisy observation vector $\boldsymbol{y}$. The elements of $\boldsymbol{y}$ corresponding to elements of mask vector $\boldsymbol{m}$ equal to 1 are the reliable coefficients $\boldsymbol{y}_r$. We use the reliable elements $\boldsymbol{y}_r$ as an approximation for the corresponding elements of $\boldsymbol{s}$, so problem (7) becomes:

$$\boldsymbol{x} = \operatorname*{argmin}_{\tilde{\boldsymbol{x}} \in \mathbb{R}^N} \{ \|\boldsymbol{A}_r\tilde{\boldsymbol{x}} - \boldsymbol{y}_r\|_2 + \lambda \|\tilde{\boldsymbol{x}}\|_1 \} \quad (8)$$

with $\boldsymbol{A}_r$ pertaining to the rows of $\boldsymbol{A}$ for which $\boldsymbol{m} = 1$. We can now use the sparse representation $\boldsymbol{x}$ obtained by solving problem (8) to estimate the clean observation vector as $\hat{\boldsymbol{s}} = \boldsymbol{A}\boldsymbol{x}$. However, since the reconstruction error will generally not be zero if we solve problem (8), we only impute the unreliable elements:

$$\hat{\boldsymbol{s}} = \begin{cases} \hat{\boldsymbol{s}}_r = \boldsymbol{y}_r \\ \hat{\boldsymbol{s}}_u = \boldsymbol{A}_u \boldsymbol{x} \end{cases} \quad (9)$$

with $\boldsymbol{A}_u$ and $\hat{\boldsymbol{s}}_u$ pertaining to the rows of $\boldsymbol{A}$ and $\hat{\boldsymbol{s}}$ for which $\boldsymbol{m} = 0$. Note that the resulting clean speech estimate $\hat{\boldsymbol{s}}$ is obtained using *unbounded* imputation: we have not taken the upper bound on clean speech estimates into account (cf. Section II-E). While *bounded* imputation would probably better be implemented by adapting the minimization problem (6) (cf. Section VIII-C), we have opted for a computationally more convenient solution, i.e., we reject those elements of which

we are sure they have been estimated incorrectly incorrectly because the estimate exceeds the observed noisy speech. For that purpose we modify (9) as follows:

$$\hat{\boldsymbol{s}} = \begin{cases} \hat{\boldsymbol{s}}_r = \boldsymbol{y}_r \\ \hat{\boldsymbol{s}}_u = \min\left(\boldsymbol{A}_u \boldsymbol{x}, \boldsymbol{y}_u \right) \end{cases} \quad (10)$$

with the $\min$ operation taking the element-wise minimum of two values.

A version of $\hat{\boldsymbol{s}}$ that is reshaped into a $K \times T$ matrix can be considered a denoised spectrogram of the underlying speech signal and can directly be used for speech decoding.

C. Minimum proportion of reliable features for successful imputation

The question arises how much missing data can be imputed using sparse imputation. Obviously, no imputation is possible if $\boldsymbol{y}$ does not contain any reliable coefficients. In practice, a minimum number of reliable coefficients will be required for successful restoration of $\boldsymbol{y}$. However, it is not possible to give an exact lower bound for the proportion of reliable features needed for successful imputation.

A necessary condition for the recovery of $\boldsymbol{x}$ is given in [26]:

$$\|\boldsymbol{x}\|_0 \lesssim \frac{F + 1}{3} \quad (11)$$

with F the number of 'measurements' of $\boldsymbol{x}$. Thus, at least $F = (\|\boldsymbol{x}\|_0 \cdot 3) - 1$ measurements (in our case, observed reliable features in $\boldsymbol{y}$) are necessary to recover $\boldsymbol{x}$. However, this does not necessarily equal the number of measurements that are sufficient to recover $\boldsymbol{x}$. Three issues play a role here.

The first issue is that, for a given speech token, we *do not know* how sparse its representation $\boldsymbol{x}$ is. While an average sparsity (i.e. the number of nonzero elements in $\boldsymbol{x}$) could be established using a representative collection of clean speech tokens, specific speech tokens may require far more or far less exemplars. Thus, any bound will depend on the individual properties of the speech token under consideration.

The second issue is that (11) is only a necessary condition. Depending on the dictionary $\boldsymbol{A}$, the real number of measurements necessary can be higher [28]. Some theoretical bounds exist (cf. [29], [30]) on the successful recovery of a sparse representation given the sparsity of $\boldsymbol{x}$ and a dictionary $\boldsymbol{A}$. Unfortunately bounds such as the Restricted Isometry Property (RIP) are sufficient, but not strictly necessary conditions and are NP-hard to establish.

The third issue is that even if we had a bound on the number of measurements needed to recover $\boldsymbol{x}$ using the dictionary $\boldsymbol{A}$, we recover $\boldsymbol{x}$ using the row-reduced dictionary $\boldsymbol{A}_r$. The Johnson-Lindenstrauss lemma [31] asserts that when points are projected onto a randomly selected subspace of suitably high dimension, the distances between the points are approximately preserved. Removing randomly selected rows from $\boldsymbol{A}$ could be considered a random mapping of $\boldsymbol{A}$ to a low dimensional version $\boldsymbol{A}_r$, thus allowing recovery of $\boldsymbol{x}$ from $\boldsymbol{A}_r$. Unfortunately, in our application the missing data is not randomly distributed. Even if the background noise was random noise, the reliable data would still be located in compact regions determined by

the speech signal (corresponding to high energy regions in the spectrogram). This makes bounds on the successful recovery of x dependent on the exact structure of A_r, which will be different from utterance to utterance.

All considerations above make it unpractical to derive bounds on successful recovery. We will therefore follow an experimental approach in which we first investigate what the sparsity is of clean speech and then try to generalize that result to noisy speech.

IV. BASELINE MISSING DATA ASR METHODS

In this Section we briefly describe two imputation methods that are among the best front-end (i.e. imputation before decoding) and best overall (employing imputation during decoding) methods in the literature on missing data techniques for ASR. The front-end method is inspired by *cluster-based imputation* [7] and is described in Section IV-A. The second method is called *per-Gaussian-conditioned imputation* [13] and is described in Section IV-B.

A. Cluster-based imputation

Consider a single time frame of the clean speech spectrogram S and the noisy speech Y and denote these by $\varsigma(t)$ and $\psi(t)$. In the cluster-based imputation front-end, we assume that every clean speech frame $\varsigma(t)$ is part of a *cluster*. Each cluster is described by a Gaussian distribution $N(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ with cluster identity $z \in Z$, mean $\boldsymbol{\mu}$ and full covariance matrix $\boldsymbol{\Sigma}$.

The cluster means are trained on a clean speech database using K-means vector quantization (VQ). Once the cluster identities of all speech frames in the database are known, we determine the covariance of each cluster.

If we know the cluster identity z of an observed noisy speech vector, its Maximum Likelihood Estimate (MLE) under the assumption of additive noise (cf. Section II-E) is:

$$\hat{\varsigma}_z = \underset{\tilde{\varsigma} \in \mathbb{R}^K}{\mathrm{argmin}} \{\frac{1}{2}(\tilde{\varsigma} - \boldsymbol{\mu}_z)'\boldsymbol{\Sigma}_z^{-1}(\tilde{\varsigma} - \boldsymbol{\mu}_z)\}$$
$$\text{subject to } \tilde{\varsigma}_u \leq \psi_u, \tilde{\varsigma}_r = \psi_r \quad (12)$$

in which we dropped the time dependency to simplify notation. The minimizer $\hat{\varsigma}_z$ is a clean speech estimate for the noisy speech frame. $\hat{\varsigma}_u$ and ψ_u denote the unreliable elements of $\hat{\varsigma}$ and ψ, respectively. Accordingly, $\hat{\varsigma}_r$ and ψ_r denote the reliable elements of $\hat{\varsigma}$ and ψ.

Since in practice we do not know the cluster identity in advance, we construct clean speech estimates $\hat{\varsigma}_z$ for all clusters Z and calculate their likelihood using:

$$f(\hat{\varsigma}_z|z) = \frac{\exp(-\frac{1}{2}(\hat{\varsigma}_z - \boldsymbol{\mu}_z)'\boldsymbol{\Sigma}_z^{-1}(\hat{\varsigma}_z - \boldsymbol{\mu}_z))}{\sqrt{2\pi}^K \sqrt{\det(\boldsymbol{\Sigma}_z)}} \quad (13)$$

Finally, we construct $\hat{s}$ as a weighted sum of cluster-conditioned clean speech estimates:

$$\hat{s} = \sum_{z=1}^{Z} \frac{f(\hat{\varsigma}_z|z)}{\sum f(\hat{\varsigma}_z|z)} \hat{\varsigma}_z \quad (14)$$

By applying this procedure for every time frame independently we obtain an estimate of a clean speech spectrogram.

B. Per-Gaussian-conditioned imputation

We used a mainstream Hidden Markov Model (HMM) based recognizer with Gaussian Mixture acoustic Models (GMM). A clean speech frame $\varsigma(t)$ is modeled by a mixture of Gaussians with diagonal covariance. We explain the imputation technique for a single Gaussian, but the results extend naturally to a mixture of Gaussians [13].

In an HMM the likelihood of observing $\varsigma(t)$ is calculated under the assumption of being in the q-th HMM state by:

$$f(\varsigma(t)|q) = N(\varsigma(t); \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \quad (15)$$

with state index q and $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ a Gaussian density function at $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and diagonal covariance $\boldsymbol{\Sigma}$.

For every Gaussian the MLE of an unreliable element is given by its corresponding Gaussian mean $\boldsymbol{\mu}$. Under the constraint of additive noise (cf. Section II-E) this gives:

$$\hat{\varsigma}_u(t) = \min(\boldsymbol{\psi}_u(t), \boldsymbol{\mu}_q) \quad (16)$$

with the min operation working element-wise.

Features in the log-spectral domain are not attractive for speech recognition because they tend to be correlated. In automatic speech recognition a linear transformation (such as for example a Discrete Cosine Transformation (DCT)) is used to decorrelate the log-spectra. Under a transformation C we express $\varsigma(t)$ as:

$$c(t) = C\varsigma(t) \quad (17)$$

with C the DCT-matrix in the case of cepstral features. Under this transformation (dropping the time dependency and index q for ease of notation) the MLE is given by:

$$\hat{\varsigma} = \underset{\tilde{\varsigma} \in \mathbb{R}^K}{\mathrm{argmin}} \{\frac{1}{2}(\tilde{\varsigma} - \boldsymbol{\mu}_\varsigma)'\boldsymbol{P}(\tilde{\varsigma} - \boldsymbol{\mu}_\varsigma)\} \text{ subject to } \tilde{\varsigma}_u \leq \psi_u \quad (18)$$

with $\boldsymbol{\mu}_\varsigma$ the Gaussian mean in the log-spectral domain. P is constructed as:

$$\boldsymbol{P} = \boldsymbol{C}'\boldsymbol{\Sigma}_C^{-1}\boldsymbol{C} + \kappa\boldsymbol{\Sigma}_S^{-1} \quad (19)$$

with $\boldsymbol{\Sigma}_C$ the diagonal covariance in the transformed domain, $\boldsymbol{\Sigma}_S$ the diagonal covariance in the log-spectral domain and κ a regularization parameter which depends on the structure of C.

The minimization problem (18) can be cast as a non-linear least squares problem and can be solved efficiently using a gradient descent or multiplicative updates method.

When modeling the speech by a mixture of Gaussians, the clean speech estimates are conditioned per-Gaussian: we get as many clean speech hypotheses for ψ as there are Gaussians in the speech model. Each Gaussian conditioned likelihood is evaluated using the imputed speech. During the Viterbi search over all likelihoods, these hypotheses are in competition with each other.

In our implementation, we did not use the cepstral transformation, but PROSPECT features (cf. [32]), a computationally efficient low order approximation of cepstral features that does not require regularization of $\boldsymbol{P}$. The speech recognizer

uses first and second time derivatives of features which are processed in a similar manner [33].

V. EXPERIMENTAL SETUP

In this Section we outline the setup of our experiments with spoken digit recognition. The recognition task is described in more detail in Section V-A. Section V-B explains the preprocessing of the speech data prior to recognition. Section V-C discusses the creation of the two types of missing data masks that are used in the experiments. The implementation of the sparse imputation algorithm and the creation of the overcomplete dictionary of exemplars are described in Section V-D. The implementation of cluster-based imputation is described in Section V-E. The speech decoder that can perform per-Gaussian-conditioned imputation is described in Section V-F.

A. Recognition task

We studied an isolated-digit recognition task using speech data from the AURORA-2 corpus [12]. The isolated-digit speech data was created by extracting individual digits from the connected digit utterances in the AURORA-2 corpus. To this end we used a segmentation obtained from a forced alignment of the clean speech utterances with the reference transcription.

The clean speech training set of AURORA-2 consists of $27\,748$ digits in 8440 utterances. The original connected digit utterances were used for extracting cluster means and covariances for cluster-based imputation (Section V-E) and for training the acoustic models of the ASR engine (Section V-F). Isolated digits extracted from these utterances were used to construct the exemplar dictionary used in sparse imputation (Section V-D).

For our experiments we used test set A, which comprises 4 clean and 24 noisy subsets. The noisy subsets are composed of four noise types (subway, car, babble, exhibition hall) artificially mixed at six SNR values, SNR= $20, 15, 10, 5, 0, -5$ dB. Every SNR subset consisted of $3257, 3308, 3353$ and 3241 digits per noise type, respectively. All experiments were carried out on the isolated, time-normalized digits.

We evaluated word recognition accuracy of the imputation methods as a function of SNR and mask type, averaging the results over the four noise types.

B. Preprocessing

Acoustic feature vectors consisted of mel frequency log power spectra: 23 frequency bands with center frequencies starting at 100 Hz (frame shift = 10 ms). All words were represented as a matrix of 35 time frames, using spline interpolation to compress longer and expand shorter word tokens. This corresponds to the average duration of the digits in the training set. Comparison with previously reported recognition accuracies of AURORA-2 clean speech (cf. [34] in which the same ASR engine was used as in the current study), shows that the time normalization does not affect recognition accuracy.

The ASR engine requires first and second time derivatives of the features. Both for cluster-based imputation and sparse

imputation these derivatives were obtained from the time-normalized representations after imputation. For per-Gaussian-conditioned imputation first and second derivatives were calculated based on the noisy (but time-normalized) spectra. Adding the derivatives results in a 69 features per frame.

C. Missing data mask estimation

The oracle mask was calculated for every digit using (2) (for AURORA-2 the power spectrograms of both clean speech S and noise N are available) with a threshold $10\log_{10}(\theta)$=-3 dB.

For the computation of the harmonicity mask, we followed the procedure described in [13]. The noisy speech signal is first decomposed in a harmonic and a residual part using a least squares fitting method. The harmonic energy can be used as an estimator of the clean speech energy and the residual as an estimator for the noise energy, for use in (2). However, the harmonic part will also contain contributions from the noise, while the residual also contains contributions from the speech. Therefore, the method uses a signal-to-noise-dependent compensation, combining harmonicity and SNR criteria. Following [13], [24] we chose $10\log_{10}(\theta)$=-9 dB. From Fig. 3 it can be seen that the harmonicity mask systematically overestimates the proportion of unreliable features (relative to the oracle mask). Experiments have shown that lowering the proportion of false unreliables raises the proportion of false reliables at at least the same rate, resulting in a lower overall recognition performance.

For per-Gaussian-conditioned imputation we calculated masks for the first and second time derivatives of features by taking derivatives of the static missing data mask (cf. [33]).

D. Sparse imputation

The sparse imputation method was implemented in MATLAB. The l^1 minimization was carried out using the `l1_ls` solver [35].[3] The regularization parameter λ was determined using the utility function `find_lambdamax_l1_ls`. The stopping criterion of the solver was a duality gap of 0.01.

A pilot study conducted to investigate the effect of the number of examples in the dictionary showed that recognition accuracy did not improve with dictionary sizes $N \geq 4000$, while computational complexity increased more than linear in the dictionary size (in [35] it was stated that the `l1_ls` solver has complexity $\mathcal{O}(N^{1.2})$). Therefore, we used a single dictionary containing 4000 exemplars that were randomly selected from the set of clean speech training exemplars. No attempt was made to represent genders, regional background or digits uniformly.

The exemplars were time-normalized in the manner described in Section V-B. Next, every digit (exemplar) was represented as a $23 \cdot 35 = 805$ dimensional vector by concatenating subsequent time-frames. The resulting $N = 4000$ exemplars were concatenated to form a single 805×4000 dimensional dictionary matrix A. Finally, the Euclidean norm of all columns were normalized to 1.

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

E. Cluster-based recognition

As in [7] we extracted means and covariances for 512 clusters from the (non time-normalized) clean speech training set of AURORA-2. First, the cluster means were calculated on $50\,000$ frames, which were randomly selected from the training set, using the `kmeans` function of the `SPIDER` toolbox.[4] Then, every frame of the $745\,761$ clean speech frames in the training set was assigned a cluster identity based on the Euclidean distance to these cluster means. Finally, we calculated for every extended cluster the new mean and covariance, resulting in 512 Gaussians of 23 dimensions with full covariance.

The bounded imputation routine was implemented in MAT-LAB and carried out using 300 multiplicative updates [36].

F. Speech recognition

For recognition we used a MATLAB implementation of the ASR engine described in [32]. This engine internally converts the spectral features to PROSPECT features (cf. Section IV-B). As in [32] we trained 11 whole-word models with 16 states per word, as well as two silence words with 1 and 3 states, respectively, using the (non time-normalized) clean speech train set of AURORA-2. Every state was modeled by 16 Gaussians with diagonal covariance.

The recognition system performs per-Gaussian-conditioned imputation during recognition, guided by a missing data mask. For the experiments with cluster-based imputation and sparse imputation we used the same recognizer, fed with clean speech estimates provided by the imputation front-ends, in combination with a mask that labels all features reliable.

VI. RESULTS

In this Section we present the results of several experiments. In Section VI-A we investigate how sparsely clean speech digits can be represented using our exemplar dictionary. We give visual examples of the output of cluster-based imputation, sparse imputation and per-Gaussian-conditioned imputation in Section VI-B. We conclude with describing the recognition results obtained by employing the three imputation methods for both mask types and report recognition accuracy as a function of SNR in Section VI-C.

A. Sparse representation of speech

We investigated the sparsity of clean (uncorrupted) speech of isolated digits in subset 1 of the AURORA-2 test database. To compare the sparsity of different digits the observation vector was normalized to a Euclidean unit norm. For every digit, we recovered its sparse representation by solving problem (7) using a dictionary of $N = 4000$ exemplars. Then, we sorted the resulting weight vector x with respect to weight. Finally, we averaged the sorted weight vectors over all 3257 digits.

The result is a cumulative weight vector which shows the average weights of sparse representations of digits ordered with respect to the largest weights of every digit. The 40 largest
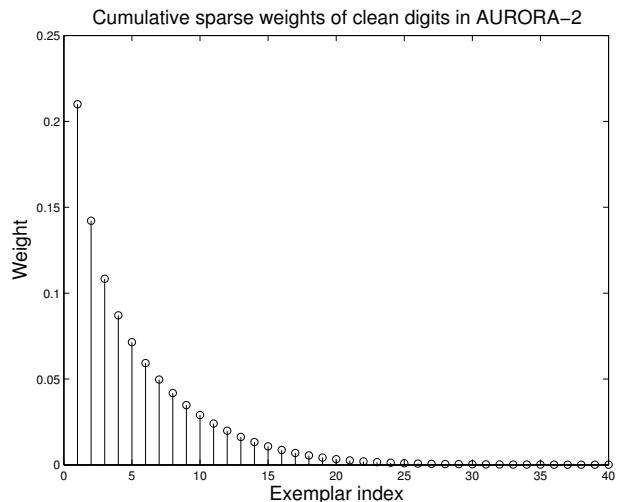

Fig. 4. The sparsity of clean speech isolated digits in subset 1 of the AURORA-2 test database. The sparse representation x of every digit is found by solving problem (7) using a dictionary of $N = 4000$ exemplars taken from the clean training database of AURORA-2. The graph shows the average weight of the 40 largest nonzero elements of each sparsely represented digit.

weights are shown in Figure 4. From this figure it can be seen that the isolated clean speech digits in the test set can indeed be sparsely represented in a dictionary of exemplar digits. The results show that there is a fast decay of the sparse weights and that on average digits can be sparsely represented using no more than approximately 25 exemplars.

B. Visual example of imputation results

In Figure 5 we show the clean speech estimates of a single isolated digit. The digit is the word "three" (pronounced /θri/ using the IPA phonetic alphabet) extracted from the utterance `MAH_1390A` which was artificially mixed with subway noise at SNR $= 5$ dB. In all cases the digit had been correctly recognized after imputation.

The clean speech estimate of per-Gaussian-conditioned imputation was created after recognition using the recognized state-sequence. This is necessary since the method creates an imputation hypothesis for every Gaussian (and thus every state). The clean speech estimate at every time frame corresponds to the imputation hypothesis of the best scoring Gaussian pertaining to the recognized state.

Comparing the clean speech spectrogram shown in Fig. 5a with the oracle mask overlayed noisy digit shown in Fig. 5d it can be seen that an imputation technique has to reconstruct the onset (the moderate energy pattern on the left of the spectrogram, indicated by ellipse number 1 in Fig. 5a) as well as the frication of the /θ/ (the high energy pattern in the upper left corner, ellipse number 2). Making the same comparison with the estimated mask overlayed noisy digit shown in Fig. 5e it can be seen that the imputation technique has to reconstruct an additional formant trace (the high energy structure in the upper right corner, ellipse number 3).

Comparing the three clean speech estimates obtained with an oracle mask of per-Gaussian-conditioned imputation, cluster-based imputation and sparse imputation shown

---

[4]The toolbox publicly is available from http://www.kyb.mpg.de/bs/people/spider/main.html

in Figs. 5f, 5g and 5h we can see substantial differences. Cluster based imputation shown in Fig. 5g clearly has retained some of the corrupting noise shown in Fig. 5c and failed to reconstruct some of the occluded high energy areas. Both per-Gaussian-conditioned imputation (Fig. 5f) and sparse imputation (Fig. 5h) have reconstructed the missing energy patterns to some extent but the clean speech estimate of per-Gaussian-conditioned imputation looks more like a checker board than the sparse imputation result.

Clean speech estimates created by cluster-based imputation employing the estimated mask shown in Fig. 5j clearly fails to reconstruct the high energy structure in the upper right corner. Per-Gaussian-conditioned imputation (Fig. 5i) and to a lesser extent sparse imputation (Fig. 5k) have succeeded in reconstructing this structure. Finally, it is worth noting that the clean speech estimates obtained using the oracle mask (Fig. 5f) and the estimated mask (Fig. 5i) are very similar when employing per-Gaussian-conditioned imputation.

C. Recognition experiments

Fig. 6 depicts the recognition accuracy on the AURORA-2 single-digit task obtained using the oracle mask. In this figure three lines are plotted corresponding to sparse imputation, per-Gaussian-conditioned imputation and cluster-based imputation. It is immediately apparent that our sparse imputation technique performs very well. While the differences between the three techniques are negligible at high SNR's (> 15 dB), sparse imputation substantially outperforms the other two imputation techniques at lower SNR's. At SNR $= -5$ dB sparse imputation obtains a recognition accuracy of 92% versus 61% for per-Gaussian-conditioned imputation and 50% for cluster-based imputation.

Fig. 7 shows the recognition accuracies of the three imputation techniques obtained with the *harmonicity* mask described in Section V-C. It can be seen that per-Gaussian-conditioned imputation now outperforms sparse imputation, while cluster-based imputation still performs worst. As with the results displayed in Fig. 6, the differences are negligible at SNR's ≥ 15 dB. Overall, the differences in accuracy between the three techniques when using the estimated (harmonicity) mask are much smaller than with the oracle mask. The largest gap between the recognition accuracies of per-Gaussian-conditioned and sparse imputation is 4.6% at SNR $= 5$ dB, while the largest difference between sparse imputation and cluster-based imputation is 8% at SNR $= 0$ dB.

VII. DISCUSSION

We first discuss the results of the experiments in Sections VII-A, VII-B and VII-C. In Section VII-D we discuss the generalizability of the findings presented in this work. Finally, we discuss related work in Section VII-E.

A. Sparse representation of speech

The experiment described in Section VI-A was carried out on clean speech, so the sparse representations were obtained using $F = D = K \cdot T = 23 \cdot 35 = 805$ measurements (features

in s). We showed that the average sparsity of clean speech digits is 25. Using the necessary condition in (11) as a best-case scenario, it can be inferred that to recover x we need at least $F = (25 \times 3) - 1 = 74$ measurements. It is unlikely that 74 reliable features of a noisy speech spectrogram are sufficient in practice, however: (un)reliable features are not randomly distributed over time and frequency and the real number of features required will depend on the dictionary A (cf. Section III-C). Still, we can use this figure to estimate a best-case upper bound on the SNR at which we can achieve 'perfect' reconstruction using the results in Fig. 3.

The 74 features amount to $74/805 \approx 9\%$ of the available features in a spectrogram. From Fig. 3 we can deduce that for the oracle mask, even at SNR $= -5$ dB on average 18% of the features is reliable, which is more than the lower bound of 9%. However, for some noisy digits the number of reliable features will be below average, leading to a erroneous imputation; this may reduce the overall recognition accuracy.

We can make an estimate of an upper bound on the SNR that still allows 'perfect' reconstruction by finding the SNR at which for most digits up to $100 - 9 = 91\%$ of the features is missing. Using the 99th percentile shown in Fig. 3 we can infer that for the oracle mask this occurs at SNR ≈ 5 dB. For the harmonicity mask the 91% limit is reached at SNR ≈ 15 dB. In other words, we can at best expect 'perfect' reconstruction for 99% of the digits for SNR's up to 5 dB for the oracle mask. Ignoring mask estimation errors we can at best expect 'perfect' reconstruction at SNR $= 15$ dB for the harmonicity mask. This is corroborated by the results in Figs. 6 and 7.

B. Visual example of imputation results

The cluster-based imputation method described in Section IV-A failed to reconstruct the high energy structures of the clean speech spectrogram outside the frames which contain reliable features both when using an oracle and an estimated mask. This is due to the frame-by-frame processing: the imputation has no knowledge of neighboring frames, neither through state-based knowledge as in per-Gaussian-conditioned imputation nor through the longer time-windows used in sparse imputation. Cluster-based imputation also retained much of the corrupting noise. This is due to the difficulty of determining cluster-identity. In our implementation we use a weighted sum of all cluster-based imputation hypotheses. While some hypotheses may contain no residual noise, the weighted sum is likely to contain residual noise due to the averaging. As noted in [7], however, choosing only one imputation hypothesis result is not a solution, due to the difficulty of selecting the proper cluster identity in the presence of noise.

Both sparse imputation and per-Gaussian-conditioned imputation succeed in reconstructing the unseen clean speech features to a large extent. In per-Gaussian-conditioned imputation this is due to the knowledge of an underlying state-sequence, in sparse imputation through the use of the large time-window.

The greater roughness of per-Gaussian-conditioned imputed spectra when compared to sparse imputation can be understood from the state/Gaussian conditioned nature. The spectra are
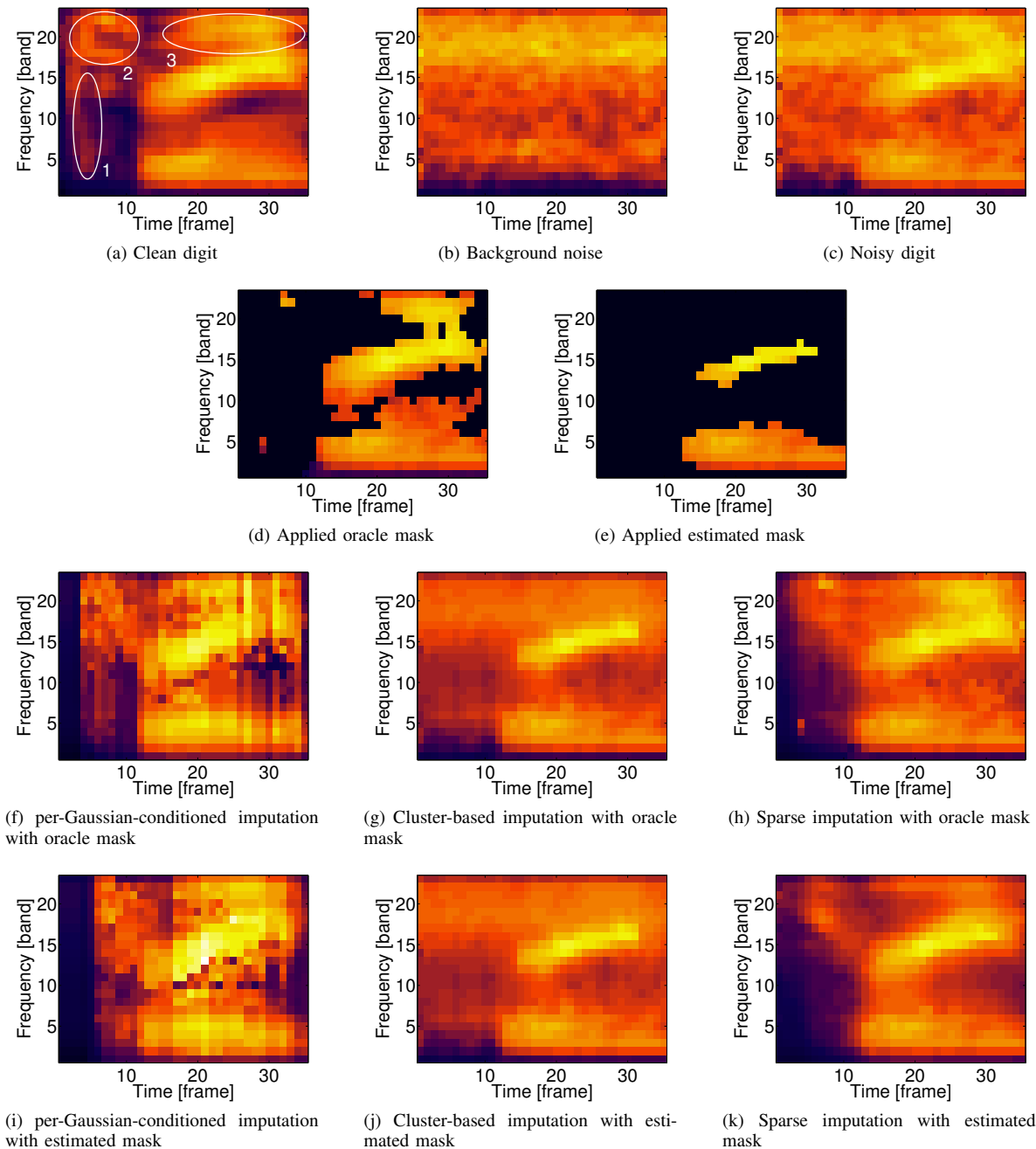
(a) Clean digit

(b) Background noise

(c) Noisy digit

(d) Applied oracle mask

(e) Applied estimated mask

(f) per-Gaussian-conditioned imputation with oracle mask

(g) Cluster-based imputation with oracle mask

(h) Sparse imputation with oracle mask

(i) per-Gaussian-conditioned imputation with estimated mask

(j) Cluster-based imputation with estimated mask

(k) Sparse imputation with estimated mask

Fig. 5.   Figure 5a shows the spectrographic representation of the digit 'three'. The horizontal axes represent time and the vertical axes frequency. The ellipses indicate areas of interest for imputation. Fig. 5b shows the spectrographic representation of the background subway noise. Fig. 5c shows the spectrographic representation of the digit artificially corrupted by the background noise at SNR = 5 dB. Figs. 5d and 5e show the noisy digit with the oracle respectively estimated mask overlayed. Figs. 5f, 5g and 5h show the imputation results of per-Gaussian-conditioned imputation, cluster-based imputation and sparse imputation respectively using the oracle mask. The imputed spectra obtained using the estimated mask are displayed in the corresponding Figs. 5i, 5j and 5k.

reconstructed based on a state-description. That means that every time a new state is entered, a different Gaussian is used for imputation. This results in the block structure in Figs. 5f and 5i with every block having a length of a few frames (recall that digits are described by 16 states in 35 time-frames).

Finally, the similarity between the per-Gaussian-conditioned reconstructed spectra employing the oracle and estimated mask is also due to its state-based nature: In both cases the digit in this example was first (correctly) recognized, after which the state sequence is used for selecting the state-dependent

clean speech estimate. Since the state sequences are very similar if the recognition result is the same, the clean speech estimates are also very similar. Consequently, when a digit is not correctly recognized, the reconstructed spectra might look very different from the clean speech spectra.

C. Recognition experiments

1) Oracle mask: The recognition accuracies displayed in Fig. 6 show that sparse imputation can successfully restore the missing data even at low SNR's. Since at SNR = −5 dB on
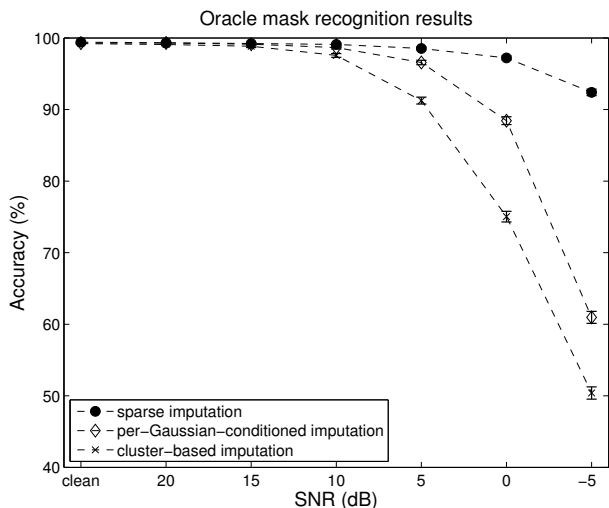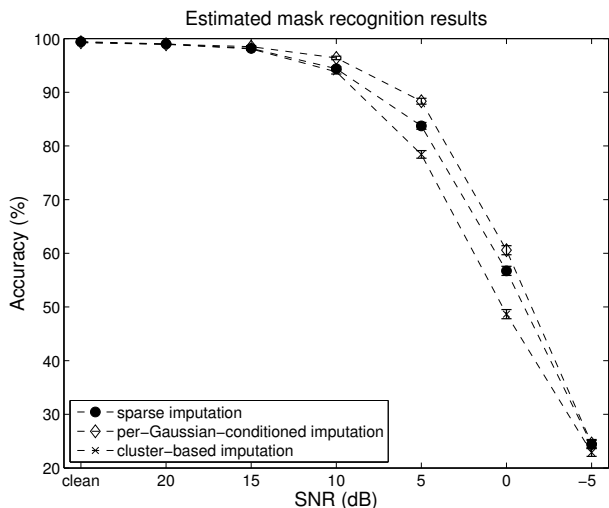
Fig. 6. Recognition results of the single digits extracted from AURORA-2. The results displayed in this figure are obtained using an oracle mask. We compare three imputation techniques: sparse imputation, per-Gaussian-conditioned imputation and cluster-based imputation. The horizontal axis describes the SNR at which the clean speech is mixed with the background noise, while the vertical axis describes recognition accuracy averaged over the four noise types described in Section V-A. The accuracy range in this figure is [40, 100]. The vertical bars around the data points indicate 95% confidence intervals.

Fig. 7. Recognition results of the single digits extracted from AURORA-2. The results displayed in this figure are obtained using an estimated mask, the *harmonicity* mask described in Section V-C. We compare three imputation techniques: sparse imputation, per-Gaussian-conditioned imputation and cluster-based imputation. The horizontal axis describes the SNR at which the clean speech is mixed with the background noise, while the vertical axis describes recognition accuracy averaged over the four noise types described in Section V-A. The accuracy range in this figure is [20, 100]. The vertical bars around the data points indicate 95% confidence intervals.

average 82% of the data is missing (cf. Fig 3), this is a very encouraging result. By contrast, recognition accuracies obtained using per-Gaussian-conditioned imputation and cluster-based imputation show a sharp decline at SNR's ≤ 5 dB. This is due to the frame-based character of these techniques: many frames contain few -if any- reliable features making successful imputation of those frames difficult.

Recognition accuracy using sparse imputation remains almost constant for SNR ≥ 10 dB. This SNR corresponds with the prediction on the basis of reliable measurements derived in Section VII-A. The decline in recognition accuracy for sparse imputation at lower SNR's can be explained by the fact that either the assumption that data is missing at random fails or because digits occasionally do not have enough reliable features.

It is interesting to note that per-Gaussian-conditioned imputation shows the steepest decline in accuracy. Because all imputation hypotheses are in competition through the Viterbi search, accuracy falls off very steeply once too many frames do not contain any reliable values.

2) Estimated mask: The recognition accuracies in Fig. 7 show a qualitatively different picture. Most strikingly, with the estimated harmonicity mask the recognition accuracies start to drop already at moderate SNR's for all three imputation methods. Also, the difference between the three methods is much smaller when compared to the oracle mask situation. Moreover, the per-Gaussian-conditioned imputation now outperforms sparse imputation.

As was the case with the oracle mask, the SNR at which the recognition performance with sparse imputation starts to break down corresponds with the prediction on the basis of reliable measurements derived in Section VII-A. However, the much steeper drop in recognition accuracies at SNR ≤ 5 dB compared to the oracle mask is somewhat unexpected. Part of the differences in recognition accuracy between harmonicity and oracle mask can be attributed to a smaller number of reliable features. The lower recognition accuracies for sparse imputation cannot entirely be explained by the reduced number of reliable features alone, however. One explanation is that mask estimation techniques suffer from two kinds of errors, unreliable features that are incorrectly labeled as reliable (false reliables) and reliable features incorrectly labeled as unreliable (false unreliables). Both errors affect imputation: false unreliables reduce the number of features we can use to recover x, while false reliables mislead the search for a correct sparse representation x. As can be inferred from Figs. 2 and 3, the harmonicity mask is tuned towards avoiding false reliables. The price to be payed, of course, is having fewer reliable elements in total.

Besides the fact that false reliables may play a role here, another factor must be taken into account: The *location* of the true reliable and unreliable features in the time-frequency plane. As was noted in [18], differences in recognition accuracy cannot be expressed simply as a function of the number of differing time-frequency cells: Some incorrectly labeled spectro-temporal elements may hardly affect recognition, while others are crucial for discriminating between different words. Apparently, the set of features that are classified as reliable by the harmonicity mask at lower SNR's contain (much) less information about the word identity compared to the oracle mask situation.

Mask estimation procedures are more likely to correctly label large coherent areas reliable because speech energy tends to be concentrated in coherent regions of the time-frequency plane. From a compressive sensing perspective this is not

ideal, because the measurements are not sampled randomly. Moreover, changes in the mask estimation algorithm, such as changing the threshold θ, are likely to yield fewer or more reliable features in the same coherent regions. These reliable features might be much less informative than a single reliable feature in a different area of the time-frequency plane.

We conclude that the harmonicity mask, already at moderate SNR's, fails to label some "crucial" features as reliable, making it impossible to correctly impute prior to decoding. Features that are the most likely to be incorrectly labeled unreliable are the low energy features in the consonant parts (like the /θ/ in the digit "three"). Yet, the consonant parts that are extremely important for discriminating between different digits that have similar vowels.

3) Per-Gaussian-conditioned imputation vs sparse imputation: An intriguing question that remains is why sparse imputation performs much better than the other imputation methods when using the oracle mask, while the per-Gaussian-conditioned imputation performs best when using an estimated mask. Our current experiments do not allow to formulate a definitive answer to this question, but several plausible explanations come to mind.

(1) A first explanation is related to the assumption that the noise is additive. We will discuss this issue in more detail in Section VIII-C. (2) It is also possible that sparse imputation is simply much more sensitive to false reliables than per-Gaussian-conditioned imputation: in per-Gaussian-condition imputation a false reliable only affects the imputation of a single frame, while neighboring frames are only indirectly affected through the Viterbi search, which takes place over all possible frame based imputations. In contrast, in sparse imputation, a single false reliable influences the search for x over multiple frames (in our case entire words). Thus, what appears to be a strength when using oracle masks −only a few reliable features are needed for successful imputation− may turn into a weakness as soon as the estimated mask contains a substantial number of false reliables. (3) Per-Gaussian-conditioned imputation does missing data imputation on static features as well as on the first and second time derivatives of the features as opposed to sparse imputation where only static features are imputed. With per-Gaussian-conditioned imputation the derivative features are imputed using separate masks. In contrast, in sparse imputation, derivative features are derived directly from the statics of the clean speech estimates solely to serve as input for the ASR engine. As a consequence, any incorrect imputation of the statics is only reinforced by these derivative features. In practice, this means that the recognizer may be confronted with vastly different derivative features than those seen during training.

D. Generalizability of findings

Our experiments using estimated masks were limited to the harmonicity mask. Moreover, we did not optimize the estimation procedure for the three different imputation methods. In fact, we kept the settings that resulted from previous optimization for per-Gaussian-conditioned imputation. It should be noted, however, that different imputation methods may require

different settings for optimal performance. Therefore, there is room for improvement of the performance of cluster-based and sparse imputation.

The mask estimation techniques reported in [37], [38] appeared to improve recognition accuracy in combination with per-Gaussian-condition imputation. It is reasonable to expect that mask estimation techniques can be developed that diminish the gap in performance between the oracle mask and the estimated mask for sparse imputation. Since sparse imputation outperforms per-Gaussian-conditioned imputation when using an oracle mask, we believe that sparse imputation is a promising alternative.

The experiments described in this paper are limited to recognition of single words extracted from one dataset (i.e.,the AURORA-2 corpus). Obviously, this raises questions about the generalizability of our findings to more general noisy speech recognition tasks. A set of experiments that are not reported in this paper suggest that our sparse imputation method can be extended beyond the realm of isolated AURORA-2 words. The sparse imputation framework presented here has also been used for noisy consonant recognition in the VCV-consonant challenge [39]. The sparse imputation results for that challenge were comparable with those obtained using other missing data approaches [40]. This suggests that the current findings can be replicated at least in other small vocabulary tasks. Furthermore, in [41] it was shown that the sparse imputation framework can also be extended from isolated word recognition to a connected digit recognition task (cf. Section VIII-E). Also in that work it was found that the sparse imputation approach substantially outperforms per-Gaussian-conditioned imputation when using oracle masks.

The extent to which our findings can be generalized to large vocabulary continuous speech recognition is still an open issue. In Section VIII-F we discuss in more detail how the complications of handling the much larger variability of the speech feature vectors in large vocabulary continuous speech could be addressed.

E. Related work

Independent of our work, the authors of [42] have applied l^1 minimization in a similar fashion to impute missing features of motion trajectories using the complete test set of trajectories as a dictionary. The differences with our work are that in our application the missing data is not randomly distributed, the location of missing data has to be estimated (and thus is error-prone) and that we use a separate dictionary of uncorrupted (clean speech) exemplars for missing data imputation.

Work in inpainting has utilized sparse (possibly overcomplete) dictionaries [43], [44]. The difference with our work is again that the location of the occlusions is known exactly and that these are often distributed more evenly over the pictures. Moreover, the amount of missing data in inpainting applications is typically much smaller.

Also, there is a substantial amount of work on source separation using sparse representations (e.g. [45]–[47]). These methods, however, have in common that they decompose the signal using models of all sources. In our case that would

amount to having a model of the clean speech as well as a model of the corrupting noise. In most speech recognition applications it is not possible to build a useful model of the noise.

In [6] the author proposed a covariance-based reconstruction method which also exploits the time-context during reconstruction. It works by modeling the spectral features as a stationary random process. Then, pairwise statistical correlations (i.e. correlations across frequency and time dimensions) are used to reconstruct missing regions. The method was found to perform well when features are missing at random, but was outperformed by bounded cluster-based imputation in a more realistic setting. The main difference with our method is that we make no assumptions about the statistical distributions of the underlying process, because we use an exemplar-based approach.

Finally, the speech fragment decoder approach [15], [48] is worth mentioning, in which a marginalization-based decoder simultaneously searches for a set of reliable speech fragments and a word sequence that best matches the target speaker, effectively performing a search over a large number of possible missing data masks. In this approach time-context is indirectly taken into account during the search.

VIII. FUTURE APPLICATION OF SPARSE IMPUTATION IN ASR

The sparse imputation method presented in this work outperforms cluster-based imputation, a state-of-the-art front-end based imputation technique. Therefore, the sparse imputation technique is promising for fields where adaptation of the speech decoder is undesirable or impossible, or for applications such as speech enhancement. The excellent oracle mask results also indicate that the sparse imputation technique might be useful in applications where the missing data mask is exactly known, such as bandwidth extension [49].

Additional research is needed to bridge the gap between the results obtained with the oracle mask and the estimated harmonicity mask. Several options could be explored to achieve this. Below, we discuss using probabilistic missing data masks (also known as *soft masks*) as a way to mitigate mask estimation errors (Section VIII-A), extension of the method to impute derivative features, just like the per-Gaussian-conditioned imputation method does (Section VIII-B), adapting the way in which the constraint posed by the fact that noise is additive is handled (Section VIII-C), and finally, the introduction of a sparse error term in the minimization problem to improve noise robustness (Section VIII-D).

For future application of sparse imputation to noise robust ASR it is imperative that the method is able to impute time-continuous speech. We sketch a possible extension to time-continuous ASR in Section VIII-E and discuss determining a suitable exemplar dictionary in Section VIII-F.

A. Soft missing data masks

In practical settings, especially at low SNRs, missing data mask estimation errors are unavoidable. Previous studies [18], [34], [50] have shown that the influence of mask estimation errors can be reduced when the binary reliability score is replaced by the probability that a spectral component is reliable: *soft* masks. Soft masks can be generated directly using the probabilistic output of machine learning techniques [18], or by the approach followed in [34], [50], e.g. by replacing the binary decision in (2) by a sigmoid function.

One possible approach to exploiting the additional information captured by soft masks is to replace (8) with a *weighted norm minimization*. In a weighted norm minimization problem, the reconstruction error of features is weighted by the probability that the feature is reliable. This allows the imputation to exploit more fully the information from the underlying speech signal, especially when the energy levels of noise and clean speech are approximately equal.

In [51] the use of soft masks in the sparse imputation framework is described and substantial improvements are reported.

B. Imputation of derivative features

Time derivatives of static features are known to improve recognition accuracy substantially in noise-free conditions. In a noisy environment, however, an increasing proportion of the static features becomes unreliable. As a consequence, no reliable derivative features can be computed whenever one of the static features involved in the computation appears to be unreliable. To avoid obfuscation of our experimental results related to this issue, the presented sparse imputation method was applied to static features only. In principle, however, it can be applied to any data that has a sparse representation. Since derivative features are linear combinations of time shifted log-spectra, it is likely that the sparse model holds equally well for this type of feature.

Hence, two alternative methods to handle this information come to mind. First, one could impute the derivative features independently of the static features. The imputed derivative features could then be offered as a separate information stream to the speech recognizer as is customary to ASR systems. As a second option, one could impute static and derivative features jointly, arguing that the sparse model holds for the static and derivative data jointly. Such an approach would have the additional advantage that the consistency between both streams is guaranteed. One might object that in the second option the derivative features comprise only dependent data that is being added. However, it is important to realize that the masks of the static and derivative features need not be the same so that the incorporation of derivative features does in fact enable to impose new constraints. Future research has to reveal to what extent derivative features can help reduce the overall number of imputation errors in actual practice.

C. Bounded imputation

Both cluster-based imputation and per-Gaussian-conditioned imputation employ *bounded imputation*: The imputation result is calculated using the constraint that the energy of the clean speech feature vector s (and thus the clean speech estimate $\hat{s}$) cannot exceed the energy in the noisy observation vector $\hat{s}$. Sparse imputation adheres to

this constraint by rejecting individual elements of the linear combinations of exemplars which exceed the observed energy. However, sparse imputation may still represent a noisy digit using exemplars of which the corresponding unreliable areas do exceed the observation energy. Since such exemplars may correspond to different digits, it is conceivable that we get better results if we take a different approach. One option would be to remove for every digit, prior to normalizing the columns of the dictionary, all exemplars from the dictionary which have energy values which exceed the corresponding observation energy. A more principled approach would be to constrain the minimization itself, changing (6) as follows:

$$ \boldsymbol{x} = \operatorname*{argmin}_{\tilde{\boldsymbol{x}} \in \mathbb{R}^N} \{ \|\tilde{\boldsymbol{x}}\|_1 \} \text{ subject to } \begin{cases} \boldsymbol{y}_r = \boldsymbol{A}_r \tilde{\boldsymbol{x}} \\ \boldsymbol{y}_u \geq \boldsymbol{A}_u \tilde{\boldsymbol{x}} \end{cases} \qquad (20) $$

The interior point technique [35] used in this work cannot be used to solve problem (20). Thus, investigating the extent to which such a formulation can improve recognition accuracy will require the use of general-purpose solvers or the development of a custom solver.

D. Error Correction

The sparse imputation method may be misled by features that are erroneously labeled as reliable by the mask estimation procedure. In [11] the authors achieve robustness against corruption in face recognition by including an error term in the minimization problem.

Assuming that most, if not all, reliable features are correctly identified by the mask estimation procedure, it is reasonable to assume that an error vector $\boldsymbol{e}$ (describing which elements of the reliable feature vector $\boldsymbol{y}_r \approx \boldsymbol{s}_r + \boldsymbol{e}$ constitute false reliables) will be sparse. Accepting the fact that mask estimation will never be flawless, it might make sense to search for a sparse solution from the dictionary in combination with a sparse error vector. Thus, we could modify (8) as follows:

$$ \boldsymbol{w} = \operatorname*{argmin}_{\tilde{\boldsymbol{w}} \in \mathbb{R}^{N+V}} \{ \|\tilde{\boldsymbol{w}}\|_1 \} \text{ subject to } \boldsymbol{y}_r = [\boldsymbol{A}_r, \boldsymbol{I}]\tilde{\boldsymbol{w}} \qquad (21) $$

with V the dimensionality of the reliable feature vector $\boldsymbol{y}_r$, $\boldsymbol{I}$ the $V \times V$ identity matrix and $\boldsymbol{w} = [\boldsymbol{x}, \boldsymbol{e}]'$ with the error $\boldsymbol{e} \in \mathbb{R}^V$. Using this formulation, errors incoherent with respect to the dictionary $\boldsymbol{A}$ will be captured by activations of the identity matrix $\boldsymbol{I}$ as encoded in $\boldsymbol{e}$. In [52] it was shown that such an approach can handle large and even dense errors effectively. Investigating to what extent such a formulation can reduce the effect of false reliables is left as future work.

E. Time-continuous imputation

The promising results obtained with sparse imputation raise the question how applicable this technique might be for applications in large vocabulary continuous speech recognition. Continuous speech recognition differs in three aspects from isolated word recognition: we do not know the word-boundaries in advance, the utterances may vary in duration so that time-normalization is no longer an option and the

intrinsic variability of the speech is much larger in a large vocabulary task. In practice, this means that we have to adapt both the exemplar dictionary (to account for the larger variability in speech and the lack of duration invariance) and the imputation technique (to deal with the continuous, non-segmented character).

Given a suitable exemplar dictionary (discussed in more detail in the next Section), one possible approach is to apply sparse imputation using a sliding time window of a fixed number of frames: imputation in every window is treated as a separate imputation problem. One can use overlapping windows to provide robustness for windows that contain few -if any- reliable elements. Overlapping windows would also result in several overlapping imputation candidates. This can be handled by using for example averaging or more elaborate schemes that take the estimated quality (confidence) of the imputation into account. While using overlapping time windows leads to an increase in computational complexity, this increase is linear in the number of overlapping windows. First experiments with this approach are presented in [41].

F. Dictionary selection

In this work, the exemplar dictionary was created by a random selection from a larger set of exemplar digits. While this approach showed promising results, it is easy to see how it could be improved. A better dictionary could result in sparser solutions (thus allowing reconstruction with fewer measurements), and provide robustness against duration variation and time-shifts in continuous speech recognition. Another issue is that in large-vocabulary continuous speech the variability of the speech feature vectors is much larger. The digits $0, 1, \cdots, 9$ do not comprise all phonemes of English, and an even smaller fraction of the diphones and triphones.

For time-continuous imputation we need an exemplar dictionary which can sparsely represent arbitrary speech. Shift-invariance can be handled algorithmically [53] or through inclusion of time-shifted variants of exemplars in the dictionary. A simple extension of our random selection method would consist of randomly selecting fixed-length time windows from continuous speech utterances in the training set. This provides shift invariance and will cover variability in duration. However, it is unlikely that such an exemplar dictionary will capture the full variance of speech with a dictionary of a few thousand exemplars. A possible way to improve the dictionary would be by clustering a much larger number of exemplars and include only a few thousand cluster centroids in the eventual dictionary.

Much work has been done on dictionary learning (e.g. [54], [55]). A substantial part of this work, however, deals with building *atomic* dictionaries: Signals are described as combinations of low(er) dimensional dictionary elements, called 'atoms'. While a clean speech signal can be sparsely described by an atomic dictionary (e.g. [45]), its sparse representation in the row-reduced dictionary (for imputation of missing data) will most likely not be equal to its sparse representation of clean speech, preventing the imputation of the missing elements. In other words: Such dictionary elements give us no information about the missing parts of the spectrogram.

IX. Conclusions

In this paper we introduced a non-parametric, exemplar-based method for reconstructing clean speech from noisy observations, based on techniques from the field of Compressive Sensing. While conventional imputation techniques for ASR employ parametric models and impute the missing data on a frame-by-frame basis, our method, dubbed *sparse imputation*, can impute missing data using larger time windows such as entire words. Using an overcomplete dictionary of clean speech exemplars, the technique first finds the sparsest combination of exemplars which jointly approximate the non-missing features of a noisy speech signal. Next, that linear combination of clean speech exemplars is used to replace the missing features.

We compared our front-end based method with two state-of-the-art baseline methods: a front-end based technique, *cluster-based imputation* and a technique in which imputation is integrated in the speech decoding, *per-Gaussian-conditioned imputation*. Our results show that sparse imputation performs much better than the two baseline methods when using an oracle mask, with a recognition accuracy of 92% at SNR $= -5$ dB. With error-prone estimated masks sparse imputation performs slightly worse than per-Gaussian-conditioned imputation, but it achieves higher accuracies than cluster-based imputation.

We have discussed ways for improving the performance of sparse imputation with estimated masks and outlined a strategy for extending the approach to large vocabulary continuous speech recognition.

References

[1] S. Ahmad and V. Tresp, "Some solutions to the missing feature problem in vision," in *Advances in Neural Information Processing Systems 5*, S. Hanson, J. Cowan, and C. Giles, Eds. Morgan Kaufmann Publishers, 1993, pp. 393 – 400.

[2] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. J. Wiley & Sons, 1973.

[3] B. Raj, R. Singh, and R. Stern, "Inference of missing spectrographic features for robust automatic speech recognition," in *Proc. International Conference on Spoken Language Processing*, 1998, pp. 1491–1494.

[4] M. Cooke, P. Green, L. Josifovksi, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[5] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[6] B. Raj, "Reconstruction of incomplete spectrograms for robust speech recognition," Ph.D. dissertation, Camegie Mellon University, 2000.

[7] B. Raj, M. Seltzer, and R. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.

[8] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[9] 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, 2006.

[10] Y. Zhang, "When is missing data recoverable?" *CAAM Technical Report TR06-15, Rice University, Houston*, 2006.

[11] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[12] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ISCA ASR2000 Workshop, Paris, France*, 2000, pp. 181–188.

[13] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. of IEEE ICASSP*, vol. 1, 2004, pp. 213–216.

[14] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic Publishers, 1996.

[15] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, pp. 5–25, 2005.

[16] A. Nadas, D. Nahamoo, and M. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 1495–1503, 1989.

[17] 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," in *Proc. of INTERSPEECH-1999*, 1999, pp. 2407–2410.

[18] M. Seltzer, B. Raj, and R. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp. 379–393, 2004.

[19] W. Kim and R. M. Stern, "Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise," in *Proc. of IEEE ICASSP*, 2006.

[20] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 58–67, 2006.

[21] N. Ma, P. Green, J. Barker, and A. Coy, "Exploiting correlogram structure for robust speech recognition with multiple speech sources," *Speech Communication*, vol. 49, no. 12, pp. 874 – 891, 2007.

[22] C. Cerisara, S. Demange, and J.-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 443–457, 2007.

[23] 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. of Eurospeech*, 2001.

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

[25] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[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, 2006.

[27] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[28] D. L. Donoho and J. Tanner, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," preprint, *http://www.math.utah.edu/∼tanner/*, 2007.

[29] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[30] E. J. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse problems*, vol. 23, no. 3, pp. 969–985, June 2007. [Online]. Available: http://www.acm.caltech.edu/∼emmanuel/papers/PartialMeasurements.pdf

[31] W. Johnson and J. Lindenstrauss, "Extensions of lipschitz embeddings into a hilbert space," *Contemporary Mathematics*, vol. 26, no. 10, pp. 189–206, 1984.

[32] H. Van hamme, "Prospect features and their application to missing data techniques for robust speech recognition," in *Proc. of INTERSPEECH-2004*, 2004, pp. 101–104.

[33] ——, "Handling time-derivative features in a missing data framework for robust automatic speech recognition," in *Proc. of IEEE ICASSP*, 2006.

[34] M. V. Segbroeck and H. V. hamme, "Robust speech recognition using missing data techniqies in the prospect domain and fuzzy masks," in *Proc. of IEEE ICASSP*, 2008, pp. 4393–4396.

[35] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l1-regularized least squares," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, dec 2007.

[36] F. Sha, L. K. Saul, and D. D. Lee, "Multiplicative updates for nonnegative quadratic programming in support vector machines," in *in Advances in Neural Information Processing Systems 15*. MIT Press, 2002, pp. 1041–1048.

[37] M. Van Segbroeck and H. Van hamme, "Vector-Quantization based mask estimation for missing data automatic speech recognition," in *Proc. ICSLP*, Antwerp, Belgium, Aug. 2007, pp. 910–913.

[38] J. F. Gemmeke, Y. Wang, M. V. Segbroeck, B. Cranen, and H. V. hamme, "Application of noise robust mdt speech recognition on the speecon and speechdat-car databases," in *Proc. of INTERSPEECH 2009*, 2009.

[39] M. Cooke and O. Scharenborg, "The interspeech 2008 consonant challenge," *Proc. of INTERSPEECH 2008*, 2008.

[40] J. Gemmeke and B. Cranen, "Noise reduction through compressed sensing," *Proc. of INTERSPEECH 2008*, 2008.

[41] ——, "Missing data imputation using compressive sensing techniques for connected digit recognition," in *Proceedings of DSP 2009*, 2009.

[42] S. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories," in *IEEE International Conference on Computer Vision and Pattern Recognition, 2008*, 2008.

[43] M. Elad, J.-L. Starck, D. Donoho, and P. Querre, "Simultaneous cartoon and texture image inpainting using morphological component analysis (mca)," *Journal on Applied and Computational Harmonic Analysis*, vol. 19, pp. 340–358, Nov 2005.

[44] M. Fadili, J.-L. Starck, and F. Murtagh, "Inpainting and zooming using sparse representations," *The Computer Journal*, vol. 52, no. 1, pp. 64–79, 2009.

[45] M. N. Schmidt and R. K. Olsson, "Linear regression on sparse features for single-channel speech separation," *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on (WASPAA)*, 2007.

[46] M. E. D. T. Blumensath, "Compressed sensing and source separation," in *International Conference on Independent Component Analysis and Signal Separation*, sept 2007.

[47] M. V. S. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete decomposition for single channel speaker separation," in *Proc. of IEEE ICASSP*, 2007, pp. 641–644.

[48] A. Coy and J. Barker, "An automatic speech recognition system based on the scene analysis account of auditory perception," *Speech Communication*, vol. 49, no. 5, pp. 384 – 401, 2007.

[49] W. Jansen and H. Van Hamme, "Prospect features and their application to missing data techniques for vocal tract length normalization," in *Proc. of INTERSPEECH-2005*, 2005, pp. 2753–2756.

[50] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP-2000*, 2000, pp. 373–376.

[51] J. Gemmeke and B. Cranen, "Sparse imputation for noise robust speech recognition using soft masks," in *Proc. of IEEE ICASSP*, 2009.

[52] J. Wright and Y. Ma, "Dense error correction via ℓ^1-minimization," submitted to *IEEE Transactions on Information Theory*, 2008, preprint: http://perception.csl.uiuc.edu/ jnwright/.

[53] M. Mœrup and M. N. Schmidt, "Shift invariant sparse coding of image and music data," *Submitted to Journal of Machine Learning Research*, 2008.

[54] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, no. 2, pp. 349–396, 2003.

[55] M. Aharon, M. Elad, and A. M. Bruckstein, "The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, November 2006.

LIST OF FIGURES