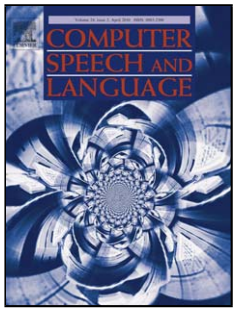
Sparse imputation for large vocabulary noise robust ASR

Jort Florent Gemmeke, Bert Cranen, Ulpu Remes



Title: Sparse imputation for large vocabulary noise robust ASR

Authors: Jort Florent Gemmeke, Bert Cranen, Ulpu Remes

PII: S0885-2308(10)00065-3
DOI: doi:10.1016/j.csl.2010.06.004
Reference: YCSLA 473

To appear in:

Received date: 29-1-2010
Revised date: 29-6-2010
Accepted date: 30-6-2010

Please cite this article as: Gemmeke, J.F., Cranen, B., Remes, U., Sparse imputation for large vocabulary noise robust ASR, *Computer Speech & Language* (2010), doi:10.1016/j.csl.2010.06.004

Sparse imputation for large vocabulary noise robust ASR

Jort Florent Gemmeke[a], Bert Cranen[a], Ulpu Remes[b]

[a] *Centre for Language and Speech Technology, Radboud University Nijmegen, P.O. Box 9103, NL-6500 HD Nijmegen, The Netherlands*
[b] *Adaptive Informatics Research Centre, Aalto University School of Science and Technology, P.O. Box 15400, FI-00076 Aalto, Finland*

Abstract

An effective way to increase noise robustness in automatic speech recognition is to label the noisy speech features as either reliable or unreliable ('missing'), and 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, frame-based imputation techniques fail because many time frames contain few, if any, reliable features. In previous work, we introduced an exemplar-based method, dubbed *sparse imputation*, which can impute missing features using reliable features from neighbouring frames. We achieved substantial gains in performance at low SNR's for a connected digit recognition task. In this work, we investigate whether the exemplar-based approach can be generalised to a large vocabulary task.

Experiments on artificially corrupted speech show that sparse imputation substantially outperforms a conventional imputation technique when the ideal 'oracle' reliability of features is used. With error-prone estimates of feature reliability, sparse imputation performance is comparable to our baseline imputation technique in the cleanest conditions, and substantially better at lower SNR's. With noisy speech recorded in realistic noise conditions, sparse imputation performs slightly worse than our baseline imputation technique in the cleanest conditions, but substantially better in the noisier conditions.

Key words: missing data techniques, noise robustness, automatic speech

Email addresses: J.Gemmeke@let.ru.nl (Jort Florent Gemmeke), B.Cranen@let.ru.nl (Bert Cranen), ulpu.remes@tkk.fi (Ulpu Remes)

recognition, sparse imputation

1. Introduction

Automatic speech recognition (ASR) performance drops rapidly when speech is corrupted with increasing levels of unfamiliar background noise (i.e., noise not seen during training) since the observed acoustic features no longer match the acoustic models. Although in real-world environments, speech is often corrupted by several unknown and time-varying noise sources, few techniques other than multi-condition training have been proposed to enhance robustness towards non-stationary noise. Missing Data Techniques (MDT) [1] are among the most promising alternative proposals.

MDT, first proposed in [2], build on the assumption that one can estimate—prior to decoding—which spectro-temporal elements in the acoustic representation of noisy speech are reliable (i.e., dominated by speech) and which are unreliable (i.e., dominated by background noise). In the unreliable elements, the clean speech information is considered *missing*, and the challenge is then to do speech recognition with partially observed data. In this work, we focus on the so-called *imputation* approach [3] which handles the missing elements by replacing them with clean speech estimates. Classic imputation methods include e.g. correlation and cluster-based reconstruction [4, 1] and methods for reconstruction in the cepstral and PROSPECT domains [5], while the state-based imputation method proposed in [6] combines front-end imputation and classifier modification.

Imputation has been proven an effective technique in both small and large vocabulary tasks [7], performing better than conventional feature-enhancement techniques (cf. [8, 9]). However, a number of issues concerning the applicability of imputation methods in different ASR tasks remain under-investigated. First, since most work on MDT has been done on artificially constructed databases (see e.g. [10, 4, 7]), the potentials and limitations of the missing data approach in real-world environments are not well known. Using artificially corrupted data is attractive as it allows estimating which features are reliable based on exact knowledge of the speech and noise power in each time-frequency cell. This facilitates comparison of different MDT approaches and allows for analysis of the influence of errors in reliability estimation. The results from such experiments are not, however, truly indicative for real-world conditions where the observed signal is rarely a simple addition of clean

speech and noise: in many cases, channel effects, the Lombard effect, and room reverberation also affect the observations.

Another issue is the imputation performance when the signal-to-noise ratio (SNR) is low and a substantial number of frames contains few, if any, reliable features. In previous work [11], we suggested that the observed performance loss when using conventional imputation methods at low SNR's is at least partly due to the fact that these methods work on a frame-by-frame basis. We argued that taking into account the time-context and utilising reliable features from neighbouring frames could reduce the number of imputation errors significantly. The use of time context for imputation has been explored in various other studies [12, 13]. Our approach to harnessing the information in neighbouring frames is by using a novel, non-parametric imputation method, *sparse imputation* (SI). We showed that the use of SI results in large performance gains and allows for successful missing data imputation at lower SNR's provided that the locations of the reliable time-frequency cells are estimated accurately [14].

The key concept in sparse imputation is that any speech fragment can be represented as a linear combination of a small number of example speech tokens. First, a dictionary of *exemplars* is constructed using fixed-length clean speech tokens. Then, a sparse linear combination of exemplars is sought using only the reliable speech features. Imputation of the unreliable features is accomplished by replacing them with the corresponding features of the linear combination of clean speech dictionary exemplars. Initially, we illustrated SI on an isolated digit recognition task where each fixed-length exemplar in the dictionary corresponded to a complete word [11]. In [15], we successfully adapted the technique for continuous digits by using a sliding window approach and a dictionary that consists of randomly selected, fixed-length segments of clean speech. During imputation, the reliable features of each window of the speech signal are treated as a sparse linear combination of clean speech windows in the dictionary. At every instant in time, the final estimates for each spectral feature vector are then calculated as an average over the reconstructions in overlapping windows.

In [15], the sparse imputation approach for continuous digits was evaluated on the AURORA-2 digit recognition task [16] which is frequently used in noise robust ASR experiments. However, it is well known that results for small vocabulary tasks are difficult to generalise to large vocabulary continuous speech recognition. One reason for this is that speech sounds occur in a much larger number of different contexts in large vocabulary tasks, which

might make it more difficult to model speech as a sparse combination of a small number of examples. This problem will only become more serious if the number of context frames in the SI approach is increased. In this paper, we will investigate to what extent the increased number of reliable features that comes from using multiple time-frames, in combination with the natural coherence of speech signals, will result in performance gains at low SNR's, despite the potential loss in accuracy due to increased variation.

In this work, we apply the sliding window approach for sparse imputation proposed in [15] on large vocabulary continuous speech data from the Finnish SPEECON database [17]. The data used in the experiments are either the original SPEECON data recorded in real-world noisy environments or artificially constructed from mixing clean speech SPEECON data and noise from the NOISEX-92 database [18]. By experimenting on different window sizes and noise types, we will investigate to what extent using more time-context can improve recognition accuracy. We will compare the results obtained with sparse imputation to results obtained with a standard frame-based parametric method, cluster-based imputation [4, 1], which has been shown to work well for the SPEECON database [19].

The rest of the paper is organised as follows. In Section 2, we discuss Missing Data Techniques for ASR and introduce the two types of reliability estimates and missing data masks used in this work. In Section 3, we briefly describe the baseline cluster-based imputation method. In Section 4, we describe the sparse imputation approach and discuss the generalisation to imputing large vocabulary speech by using a sliding time window. In Section 5, we present the experimental setup, while the results appear in Section 6 and are discussed in Section 7. Conclusions and suggestions for future research are given in Section 8.

2. Missing Data Techniques in ASR

2.1. Motivation

In this section, we briefly discuss the MDT framework as used for noise robust ASR [20, 21]. In ASR, the basic representation of speech is 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 each cell represents a combination of speech and background noise power.

Assuming noise is additive, the power spectrogram of noisy speech can be approximately described as the sum of the individual power spectrograms of clean speech and noise. To mimic human hearing, often a Mel-frequency scale and logarithmic compression of the power scale are employed. Since the logarithmic compression of a sum can be approximated by the logarithm of the largest of the two terms [22], it approximately holds for noisy speech features that:

$$\boldsymbol{Y} \approx \max(\boldsymbol{S}, \boldsymbol{N}) \tag{1}$$

with the (Mel-frequency) log-power spectrograms $\boldsymbol{Y}$ denoting noisy speech, $\boldsymbol{S}$ denoting clean speech and $\boldsymbol{N}$ representing the background noise. From (1) we can infer that the noisy speech features dominated by clean speech energy remain approximately uncorrupted and can be used directly as estimates of the clean speech features. The noise dominated features, on the other hand, provide only an upper bound for the clean speech, which means the clean speech features cannot be observed and are effectively missing.

2.2. Missing data masks

Elements of $\boldsymbol{Y}$ that predominantly contain speech or noise energy are distinguished by introducing a spectrographic mask $\boldsymbol{M}$. The elements of a mask $\boldsymbol{M}$ are either 1, meaning that the corresponding element of $\boldsymbol{Y}$ is dominated by speech ('reliable') or 0, meaning that it is dominated by noise ('unreliable' c.q. 'missing'). Thus, we write:

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

with $\boldsymbol{M}$, $\boldsymbol{Y}$, $\boldsymbol{S}$, and $\boldsymbol{N}$ two-dimensional matrices of $K \times T$, with frequency-band index k, $1 \leq k \leq K$ and time-frame index t, $1 \leq t \leq T$. θ denotes a constant SNR-threshold.

Smaller values of θ will result in more elements being considered as reliable in the mask but the proportion of errors implied in the assumption that $\boldsymbol{S}(k,t) = \boldsymbol{Y}(k,t)$ will be larger. Larger values of θ lead to a safer model but also to fewer reliable elements for estimating the missing values.

2.3. Estimating missing data masks

In experiments with artificially added noise, an *oracle mask* can be computed directly with (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 [23, 19], machine learning approaches to mask estimation [24, 25, 26], methods that focus on speech characteristics, e.g. harmonicity based SNR estimation [8, 27], and mask estimation exploiting binaural cues [28] or correlogram structure [29] (cf. [30] and the references therein for a more complete overview of mask estimation techniques).

In this work, we use the mask estimation approach described in [19]. Local SNR's are obtained from comparing the noisy speech to a noise estimate which is calculated based on frames identified as non-speech by a speech/non-speech classifier. Implementation details are given in Section 5.4. Since imputation accuracy is dependent on the quality of the missing data mask, we investigate the influence of mask estimation errors by additionally using oracle masks when the test set contains speech artificially corrupted with background noise.

2.4. Use of MDT in ASR

Techniques for speech recognition with missing data can be divided in two categories: marginalisation and imputation. In the marginalisation approach [10, 31], 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 [12, 4], the missing features are replaced by clean speech estimates, after which recognition can proceed without modification of the recognition system.

The marginalisation approach has been shown to be more robust against data sparsity at low SNR's than the traditional imputation methods [10]. Imputation methods are, however, attractive for two reasons. First, after the missing features have been replaced with clean speech estimates, any recogniser developed for clean speech can be deployed without further modifications. Another benefit is that the reconstructed features can be converted to an arbitrary domain, like the cepstral domain. This is advantageous since cepstral features are known to be less correlated and better suited for processing with the state-of-the-art HMM-based ASR techniques [32]. Therefore, in this work, only imputation techniques are investigated.

3. Cluster-based imputation

As mentioned in the introduction, we use the cluster-based imputation method proposed in [4] as the baseline approach for missing data speech recognition. It is a frame-based method where the unreliable feature values are estimated based on information in the observed features and a parametric clean speech model.

3.1. Modelling assumptions

The clean speech distribution model used in cluster-based imputation [4, 1] assumes the clean speech vectors $\boldsymbol{s}(t)$ are independent and identically distributed (*i.i.d.*). Thus, the model will capture the statistical dependencies between spectral channels but not between time-frames. It is also assumed the clean speech data can be clustered so that the features in each cluster are approximately normally distributed, and the clean speech can be modelled using a Gaussian mixture model (GMM):

$$P(\boldsymbol{s}(t)) = \sum_{\nu} P(z(t) = \nu) N[\boldsymbol{s}(t); \boldsymbol{\mu}(\nu), \boldsymbol{\Sigma}(\nu)] \; \forall \; t, \tag{3}$$

where ν are the cluster indices, $z(t)$ indicates the current cluster, $P(z(t) = \nu)$ are the cluster weights i.e. prior probabilities for $z(t)$, and $\boldsymbol{\mu}(\nu)$ are the cluster means and $\boldsymbol{\Sigma}(\nu)$ the covariance matrices. Here, the cluster identities $z(t)$ underlying $\boldsymbol{s}(t)$ are assumed unknown and modelled as a latent variable. In this work, the clusters and the distribution parameters $\boldsymbol{\mu}(\nu)$ and $\boldsymbol{\Sigma}(\nu)$ were jointly estimated from a clean speech training corpus using the expectation-maximisation (EM) algorithm.

3.2. Missing data imputation

The noisy observations $\boldsymbol{y}(t)$ corresponding to individual frames of the spectrogram $\boldsymbol{Y}$ are divided into mutually exclusive reliable and unreliable regions $\boldsymbol{y}_r(t)$ and $\boldsymbol{y}_u(t)$ as indicated by the missing data mask (2). In cluster-based imputation [4, 1], the clean speech estimates or *reconstructions* for the unreliable features are chosen so that 1) the reconstructed vectors $\hat{\boldsymbol{s}}(t) = \boldsymbol{s}_r(t) \cup \hat{\boldsymbol{s}}_u(t)$ are similar to clean speech i.e. provide the best possible fit with the clean speech distribution model while 2) the reconstructed values $\hat{\boldsymbol{s}}_u(t)$ are constrained not to exceed the observed values $\boldsymbol{y}_u(t)$. Finding such reconstruction can be written as a bounded maximum a posteriori (BMAP)

estimation task, where the BMAP estimator for the unreliable features is given as

$$\hat{\boldsymbol{s}}_u = \underset{\boldsymbol{s}_u \in \mathbb{R}^U}{\operatorname{argmax}} \{P(\boldsymbol{s}_u | \boldsymbol{s}_r, \boldsymbol{s}_u \leq \boldsymbol{y}_u, \Lambda)\}, \tag{4}$$

where Λ are the parameters estimated for the GMM (3) and U is the number of unreliable features in $\boldsymbol{y}(t)$. We dropped the explicit notation to indicate the dependency on t, i.e., $\boldsymbol{s} = \boldsymbol{s}(t)$ and $\boldsymbol{y} = \boldsymbol{y}(t)$. Note that the model contains two unknown variables: in addition to the unreliable feature values, the cluster identities $z = z(t)$ are unknown. In (4), the latent variable z has been marginalised, but the dependency can be made explicit and (4) written as

$$\hat{\boldsymbol{s}}_u = \underset{\boldsymbol{s}_u \in \mathbb{R}^U}{\operatorname{argmax}} \{\sum_{\nu} P(z = \nu | \boldsymbol{s}_r, \boldsymbol{s}_u \leq \boldsymbol{y}_u, \Lambda) P(\boldsymbol{s}_u | \boldsymbol{s}_r, \boldsymbol{s}_u \leq \boldsymbol{y}_u, \Lambda, \nu)\}, \tag{5}$$

where the first probability term is the posterior probability for the ν-th GMM cluster given the reliable features $\boldsymbol{y}_r(t)$ and the upper bound given by the unreliable features $\boldsymbol{y}_u(t)$, and the second is the cluster-conditional posterior probability for the unreliable features.

In practice, finding maximum a posteriori estimates for GMM-distributed variables is difficult. Therefore, in cluster-based imputation [4, 1], (5) is approximated as

$$\hat{\boldsymbol{s}}_u = \sum_{\nu} P(z = \nu | \boldsymbol{s}_r, \boldsymbol{s}_u \leq \boldsymbol{y}_u, \Lambda) \underset{\boldsymbol{s}_u \in \mathbb{R}^U}{\operatorname{argmax}} \{P(\boldsymbol{s}_u | \boldsymbol{s}_r, \boldsymbol{s}_u \leq \boldsymbol{y}_u, \Lambda, \nu)\}, \tag{6}$$

where the latter term is the cluster-conditional BMAP estimate for the unreliable features $\boldsymbol{s}_u$. The cluster-conditional estimate for $\boldsymbol{s}_u$ is weighted with the posterior probability for cluster ν which is calculated based on the prior probability $P(z(t) = \nu)$ and cluster-conditional observation probability. In this work, we use full covariance matrices $\boldsymbol{\Sigma}(\nu)$ and calculate the cluster-conditional BMAP estimates iteratively over the frequency channels k as proposed in [4, 1]. The covariance matrices are only assumed diagonal when evaluating the posterior probabilities for $z(t) = \nu$.

4. Sparse Imputation

In sparse imputation, speech tokens are represented as a linear combination of tokens from an overcomplete dictionary of noise-free exemplars

represented as fixed-length vectors. For an unknown speech token, a sparse linear combination is sought in the dictionary using all reliable features in the entire token. Imputation of the unreliable features is then accomplished by replacing them with the corresponding values from this linear combination of the clean speech dictionary exemplars. In [11], the tokens were chosen to constitute time-normalised complete words, but for the continuous large vocabulary speech used in this work, we must apply the sliding window approach proposed in [15].

4.1. Sparse representation of speech

The log-power spectrogram of clean speech, $\boldsymbol{S}$, is reshaped to a single vector $\boldsymbol{s}$ of dimension $D = K \cdot T$ by concatenating T subsequent K-dimensional time frames. For now, we assume that T is fixed. Inspired by a similar approach in the field of face recognition [33], we assume that $\boldsymbol{s}$ can be represented exactly (or at least approximated with sufficient accuracy) by a linear, non-negative, combination of exemplar spectrograms $\boldsymbol{a}_n$, where n denotes a specific exemplar ($1 \leq n \leq N$) in the dictionary of N available exemplars:

$$\boldsymbol{s} = \sum_{n=1}^{N} x_n \boldsymbol{a}_n = \boldsymbol{A}\boldsymbol{x} \quad \text{subject to} \quad \boldsymbol{x} \geq 0 \tag{7}$$

with $\boldsymbol{x}$ an N-dimensional activation vector. The matrix $\boldsymbol{A}$ denotes an overcomplete dictionary: $\boldsymbol{A} = [\boldsymbol{a}_1 \ \boldsymbol{a}_2 \ldots \boldsymbol{a}_N]$, with dimensions $D \times N$ with $N \gg D$. A schematic representation of this process for a set of non-noisy spoken digits (1 through 9, "zero", and "oh") is displayed in Fig. 1 A.

Although the system of linear equations in (7) has no unique solution, research in the field of Compressive Sensing [34, 35, 36] has shown that under mild conditions on $\boldsymbol{A}$, the activation vector $\boldsymbol{x}$ can be *uniquely* determined if $\boldsymbol{x}$ is sufficiently *sparse*. This is accomplished by solving:

$$\boldsymbol{x} = \underset{\tilde{\boldsymbol{x}} \in \mathbb{R}^N}{\operatorname{argmin}} \{ \, \|\boldsymbol{A}\tilde{\boldsymbol{x}} - \boldsymbol{s}\|_2 + \lambda \|\tilde{\boldsymbol{x}}\|_1 \, \} \tag{8}$$

with a regularisation parameter λ. The requirement that the linear combination must be sparse means that it must be possible to represent speech tokens with a small number of exemplars, resulting in a small number of nonzero values in $\boldsymbol{x}$. For spoken digits, it was shown in [14] that the representation is indeed sparse.

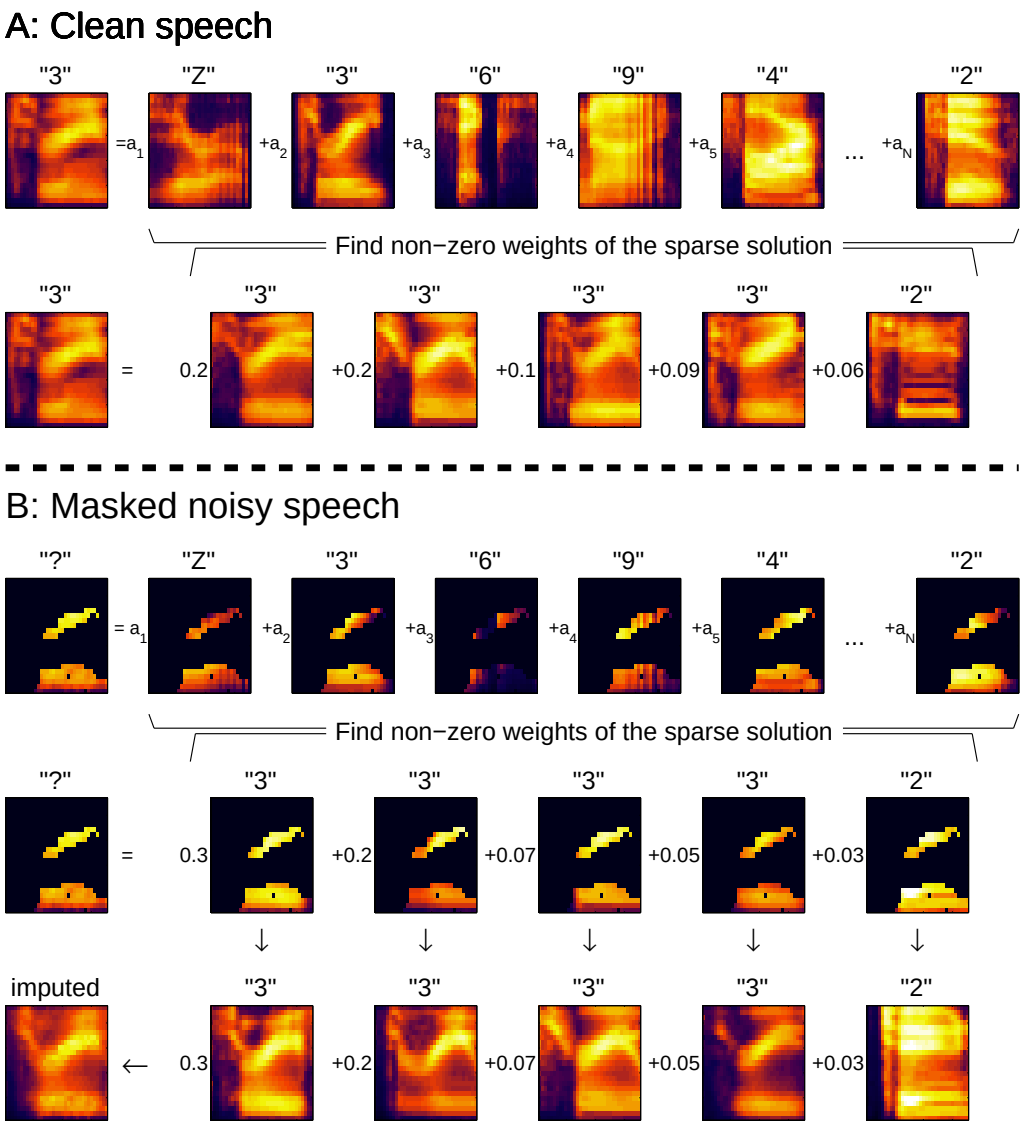


Figure 1: Schematic representation of sparse imputation, using isolated, pre-segmented digits as an example. Digit labels are at the top of the log-power spectrograms, where "Z" denotes "zero" and "O" denotes "oh". Panel A shows the sparse representation of the digit "three" in the case of clean, unmasked speech. Panel B shows the same digit with background noise at −5 dB SNR. The missing data (in black) is replaced by the corresponding features of the linear combination of clean speech dictionary exemplars found. In both panels, only the five largest nonzero weights of the linear combination are shown.

4.2. Missing data imputation

If the data contains missing values, we begin by concatenating subsequent time frames of the spectrographic mask $\boldsymbol{M}$ discussed in Section 2.2 to form a mask vector $\boldsymbol{m}$; this is done similarly as described for $\boldsymbol{s}$ in the previous section. 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 the mask vector $\boldsymbol{m}$ that are 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 (8) becomes:

$$\boldsymbol{x} = \underset{\tilde{\boldsymbol{x}} \in \mathbb{R}^N}{\operatorname{argmin}} \{ \, \|\boldsymbol{A}_r \tilde{\boldsymbol{x}} - \boldsymbol{y}_r\|_2 + \lambda \|\tilde{\boldsymbol{x}}\|_1 \, \} \tag{9}$$

with $\boldsymbol{A}_r$ pertaining to the rows of $\boldsymbol{A}$ for which $\boldsymbol{m} = 1$. The sparse representation $\boldsymbol{x}$ obtained by solving problem (9) could be directly used to estimate the clean observation vector as $\hat{\boldsymbol{s}} = \boldsymbol{A}\boldsymbol{x}$. A schematic representation of this process for a set of pre-segmented spoken digits (1 through 9, "zero", and "oh") is displayed in Fig. 1 B.

In practice, the sparse representation is not directly used as a clean speech estimate since the reconstruction error for the reliable coefficients will generally be non-zero if we solve problem (9), so it is better to only impute the unreliable elements. Furthermore, under the assumption that noise and speech are additive in the power domain, the observed noisy speech $\boldsymbol{y}$ is an upper limit for $\hat{\boldsymbol{s}}$. Incorporating these two modifications we obtain:

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

with $\boldsymbol{A}_u$ and $\hat{\boldsymbol{s}}_u$ pertaining to the rows of $\boldsymbol{A}$ and $\hat{\boldsymbol{s}}$ for which $\boldsymbol{m} = 0$ and with the *min*-operator taking the element-wise minimum of two values. A version of $\hat{\boldsymbol{s}}$ that is reshaped into a $K \times T$ matrix $\hat{\boldsymbol{S}}$ can be considered a denoised spectrogram representing the underlying speech signal, and as illustrated in Fig. 1 B, it can be directly used in speech recognition.

4.3. Sliding window approach

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 [11]. Since arbitrary length utterances clearly do not satisfy this constraint, we adopt a sliding window approach introduced in [15]. In this approach,

each window is imputed separately using sparse imputation as described in Section 4.2. Subsequently, at every time frame, the different clean speech estimates resulting from any overlapping windows are combined.

Consider a noisy speech utterance $\boldsymbol{Y}_{tot}$ represented as a spectrogram with K frequency bands and T_{tot} time-frames. The goal of the missing data imputation process is to provide an estimate $\hat{\boldsymbol{S}}_{tot}$ of the underlying clean speech $\boldsymbol{S}_{tot}$ using a missing data mask $\boldsymbol{M}_{tot}$.

We slide a window of length T_w through $\boldsymbol{Y}_{tot}$, with shifts of Δ,$1 \leq \Delta \leq T_w$ frames (cf. Fig. 2). $\boldsymbol{Y}_w$ and $\boldsymbol{M}_w$ describe the noisy speech and associated missing data mask for each window w, $1 \leq w \leq W$. The ratio of Δ and T_w determines the degree with which subsequent windows overlap. Larger step sizes Δ reduce computational effort, but can decrease imputation accuracy [15]. Throughout this paper, we keep the window shift constant at $\Delta = 1$ frame.The total number of windows we process is $W = T_{tot} - T_w + 1$.

We then use, for each window, the sparse imputation approach described in Section 4.2 to provide a clean speech estimate $\hat{\boldsymbol{S}}_w$ of the underlying clean speech $\boldsymbol{S}_w$. Since windows overlap, each frame in $\boldsymbol{Y}$ is associated with multiple clean speech estimate candidates, with the number of candidates ranging from 1 (at the beginning and end of an utterance) to T_w. For each frame, the feature values of the final clean speech estimate $\hat{\boldsymbol{S}}_{tot}$ are created by averaging over the multiple clean speech estimate candidates pertaining to that frame (cf. Fig. 2). The clean speech estimate is calculated using only clean speech estimates derived from windows with a nonzero number of reliable elements.

However, in very noisy conditions and particularly at the start and end of an utterance, it may happen that many adjacent windows do not contain reliable features, leaving the method unable to provide a clean speech estimate from averaging. Yet, despite the lack of information about the underlying signal, input must be provided to the ASR engine. Thus, we opted to impute silence (the average feature values per frequency band for silence states as observed in a training database) for such frames.

5. Experimental setup

5.1. Speech recognition system and performance evaluation

The speech recognition system used in this work is the large vocabulary continuous speech recognition system developed in the Adaptive Informatics Research Centre at the Aalto University School of Science and Technology. The acoustic models are trained with 30-hours of clean speech recorded with a

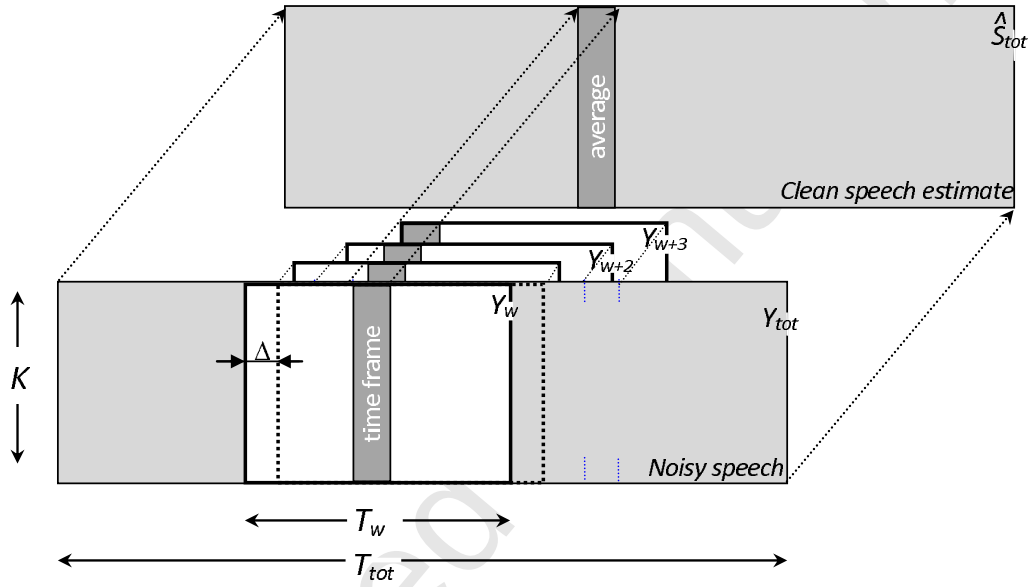


Figure 2: Schematic diagram of the sliding window approach for imputation. The dark shaded time-frame in $\boldsymbol{Y}_{tot}$ is processed in several fixed-length imputation windows, of which we have shown $\boldsymbol{Y}_{w}$ through $\boldsymbol{Y}_{w+3}$. Within each window, the given frame takes a different position due to the window shift Δ. The corresponding time-frame in the clean speech estimate $\hat{\boldsymbol{S}}_{tot}$ is the average over these individual window-based imputations.

headset in quiet conditions and selected from the Finnish SPEECON database (recorded with a 16 kHz sampling rate) [17]. The training set comprises 293 speakers (142 female and 151 male). The utterances used for training contain words, read sentences and spontaneous speech in order to have a general acoustic model valid for multiple tasks.

The decoder used in the system is a time-synchronous beam-pruned Viterbi token-pass system described in [37] and the acoustic models are state-clustered, hidden Markov triphone models, constructed with a decision-tree method [38]. There are acoustic models for 13 250 triphones and two silences. The triphones are modelled as left-right HMM with three states and silences with one state each. In total we used 1564 individual states modelled with approximately 28 Gaussians per state. Each state is also associated with gamma probability functions to model the state durations [39].

The language model employs morpheme-like subword units, called *statistical morphs*, discovered in an unsupervised, data-driven manner [40]. These are used because word-based modelling is not feasible for highly inflected languages such as Finnish, Estonian, or Turkish. The statistical morph lexicon with 25k morpheme-like units was learned from the 160k most common words extracted from 145 million words of Finnish book and newspaper data [41]. The variable-length, growing n-gram language model [42] used in this work was trained on the same text corpus and contains 52 million n-grams. The decoding vocabulary is in practice unlimited since all words and word forms can be represented using the statistical morphs [43].

Finally, in this work, the speech recognition performance is measured primarily in letter error rates (LER). This is because the words in Finnish are often long and consist of several morphemes so that measuring the word error rate (WER) would correspond better to measuring sentence or phrase error rates in languages such as English. Using the word error rate is also considered to over-penalise misrecognised word breaks.

5.2. Recognition task

The imputation methods are evaluated with clean speech recordings artificially corrupted with noise at different SNR's as well as with speech recorded in real-world noisy environments. In both conditions, the speech data consists of read sentences selected from the Finnish SPEECON database [17]. The artificially corrupted clean speech was constructed by mixing headset-recorded clean speech utterances with a randomly selected sample of the babble noise from the NOISEX-92 database [18] at SNR's 15 dB, 10 dB, 5 dB and 0 dB.

The real-world noisy speech data is recorded in two types of environments: in a car and in public places both indoors and outdoors. These recordings have been made with three microphones: 1) with a headset, 2) with a lavalier microphone, and 3) with a microphone from 0.5 m–1 m distance (in public environments) or with a microphone mounted on the rear-view mirror (in car environments).

In the SPEECON documentation, the average SNR's in the public environments are estimated to be 24 dB for the headset microphone data, 14 dB for the lavalier microphone data, and 9 dB for the far-field microphone data. For the car recordings, the estimated average SNR's are 13 dB for the headset microphone data, 5 dB for the lavalier microphone data, and 8 dB for the rear-view mirror (RVM) microphone data. The RVM microphone data has a higher SNR than the lavalier microphone data because the RVM microphone (AKG Q400 Mk3T) has a limited frequency response, specifically designed for in-car use and suppressing low frequency noise.

The speech material in each of the three scenarios (public environments, car environments, and artificially added babble noise) is divided in development and evaluation sets. The composition of the sets in terms of number of utterances (#u), duration (d), number of speakers (#s), number of female speakers (#f) and number of male speakers (#m) is shown in Table 1. None of the sets share speakers with each other or with the speech data used for training the speech recognition system.

Table 1: Composition of development and test set.

	public					car					babble				
	#u	d (min)	#s	#f	#m	#u	d (min)	#s	#f	#m	#u	d (min)	#s	#f	#m
Development	587	60	20	7	13	288	29	10	2	8	1093	115	40	22	18
Evaluation	878	94	30	13	17	575	57	20	12	8	1118	113	40	21	19

5.3. Feature extraction

Feature extraction was carried out using a 16 ms Hamming window with 8 ms overlap between subsequent windows. First-order pre-emphasis was applied to the signal using a coefficient of 0.97. After Fourier transformation, the log-power was computed in 21 triangular-shaped Mel-frequency bands. Imputation was carried out on these log-power Mel-frequency spectra to obtain clean speech estimates.

After imputation, the resulting spectra were transformed to 12 Mel-frequency cepstral coefficients (MFCC) and a log-energy feature, augmented with first and second-order time derivatives for a total of 39 features per time frame. Channel normalisation was applied using cepstral mean subtraction, and as a final step, a maximum likelihood linear transformation (MLLT). The MLLT, optimised during training of the acoustic models, is applied on the normalised features to improve the modelling of any remaining correlation in the normalised MFCCs as proposed in [44].

5.4. Missing data mask estimation

In this work, the missing data masks are constructed based on local SNR estimates obtained from comparing the noisy speech to a static noise estimate calculated during speech pauses. These speech pauses were detected using an HMM-based speech/non-speech classifier described in [19]. Additionally, we used the MATLAB command `bwareaopen` to discard small, isolated regions of reliable features from the estimated mask since it was suggested in [45] that such *glimpses* are not detectable to human listeners and are therefore unlikely to contain usable information. Experiments on the SPEECON development data also confirmed that removing glimpses comprising less than five spectro-temporal components improves speech recognition results. In experiments with artificially corrupted speech, we also computed oracle masks (Section 2.3) from which glimpses were not removed.

The SNR threshold θ for deciding whether a time-frequency component is treated as reliable or unreliable was determined by maximising recognition accuracy on the development sets described in Section 5.2. For both real-world noisy speech sets, the optimum value for estimated masks was at $\theta = 3$ dB for both imputation methods. For artificially corrupted speech, we determined an SNR-independent threshold using the development data sets containing noise at 10 and 5 dB SNR. The optimum mask threshold value for the estimated masks was at $\theta = 4$ dB for both imputation methods and for the oracle masks at $\theta = -2$ dB for sparse imputation and $\theta = -1$ dB for cluster-based imputation.

5.5. Cluster-based imputation

The clean speech model used in this work is a 5-component GMM trained using a 52-minute dataset of 500 read sentences randomly selected from the SPEECON training data described in Section 5.1. The clusters and distribution parameters are jointly estimated using the expectation-maximisation

(EM) algorithm implemented in the `GMMBayes` toolbox[1]. The cluster-conditional bounded maximum a posteriori (BMAP) estimates are calculated in an iterative manner as described in [4] using a MATLAB implementation. The estimates are calculated in 10 iterations over the mel-frequency bands.

The number of 5 clusters was selected as a reasonable balance between imputation performance and computational complexity. Non-exhaustive tests on the development data showed that while more clusters do improve recognition accuracy, the increase is slight compared to the extra computational effort required.

5.6. Sparse imputation

The sparse imputation was implemented in MATLAB. The l^1 minimisation (9) was carried out using the `SolveLasso` solver.[2] Due to instability issues, the slower, but more robust `l1_ls_nonneg` solver[3] [46] was used whenever the (fast) solver `SolveLasso` appeared to crash. When using the `l1_ls_nonneg` solver, the regularisation parameter λ was determined using the utility function `find_lambdamax_l1_ls_nonneg`. The stopping criterion of the `l1_ls_nonneg` solver was a duality gap of 0.01. The `SolveLasso` solver was run for 30 iterations.

For each window size $T_w \in \{1, 5, 10, 15, 20, 25, 30, 35\}$ being considered, an initial dictionary is created by randomly extracting 4 spectrogram segments of the desired duration T_w from each of the 8139 read sentences (containing 14 hours of speech) in the SPEECON training data described in Section 5.1. From this initial dictionary spanning 32 556 exemplars, we then randomly extract 8000 exemplars to form the final dictionary used for imputation. The dictionary size of 8000 exemplars was chosen because pilot tests showed that while using larger dictionaries improves recognition accuracy, the increase is slight enough to consider it a reasonable balance between recognition performance and computational complexity. In exemplar selection, no effort was made to balance a possible over-representation of spectra containing silence.

After reshaping the spectrograms to one-dimensional vectors as described in Section 4.1, the feature vectors form the columns of the dictionary matrix. The zero-dB level in the spectra is set such that the lowest occurring feature

[1] This toolbox is publicly available from `www2.it.lut.fi/project/gmmbayes/`

[2] This solver is implemented as part of the `SparseLab` toolbox which is publicly available from `http://www.sparselab.stanford.edu`

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

value in the dictionary is zero. Finally, the columns of the dictionary are normalised to Euclidean unit norm.

6. Results

6.1. Public and car environment data

The speech recognition results from our experiments with the data recorded in the public and car real-world environments are displayed in Fig. 3. The results depict the letter error rate (LER) of the sparse imputation (SI) method as a function of window size ($T_w \in \{1, 5, 10, 15, 20, 25, 30, 35\}$). The figure also shows the performance of the cluster-based imputation (CI) method and the baseline (B) recogniser. The latter has no noise compensation other than what is implicit in the feature extraction (see Section 5.3).

One point of interest is the fact that the LERs for SI with window length $T_w = 1$ are much higher than for any other window length. In fact, in many cases the performance at $T_w = 1$ is even worse than the baseline performance. Possible causes for this effect will be discussed in Section 7. In the rest of this section, we will largely ignore the data points at $T_w = 1$, and focus on a comparison of CI and B with SI at window lengths $T_w \geq 5$.

Another remark that holds for virtually all testing conditions is that with SI the performance seems to have an optimum in the range $T_w = [5, 20]$. Generally speaking, however, the differences obtained with various window sizes are quite small. Taking window size into account when comparing the SI results with those of the other methods would unnecessarily complicate matters. In the description below we will therefore focus on gross effects that can be observed in the range $T_w = [5, 20]$.

The first row in Fig. 3 illustrates results on the headset recorded data, the condition which resembles clean speech the most. Although some of the observed differences are statistically significant, the differences in performance of CI, SI, and B are quite small.

In the case of the lavalier microphone data, we can observe more differences between the car and public environments. In the car environment, both the baseline and CI method achieve 7 to 11 % absolute lower accuracies than on data recorded in the public environment. In the public environment, SI performs comparable with CI for window lengths in the range $T_w = [5, 20]$, while in the car environment SI outperforms CI by some 4 % (absolute).

In the case of the far-field microphone (public environment) or rear-view mirror (RVM) microphone (car environment), the results are similar as for

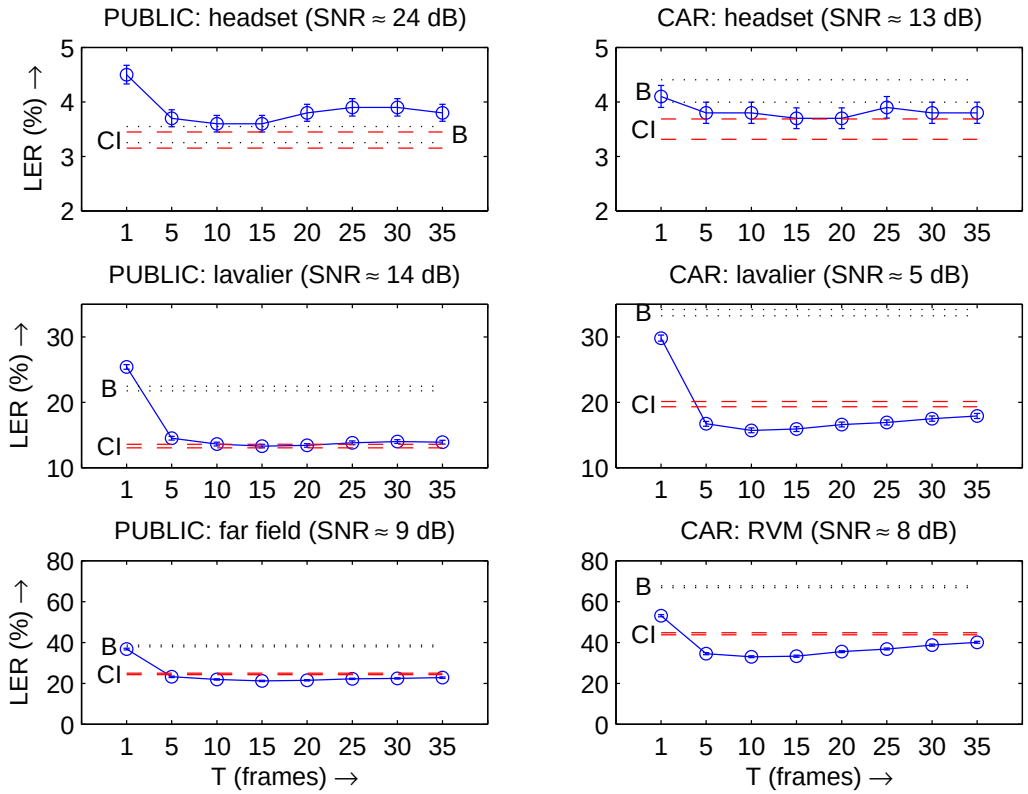


Figure 3: Recognition accuracy expressed as letter error rates (LER) for the public (left pane) and the car environment (right pane). From top to bottom, the rows correspond to headset, lavalier, and far-field microphone (public environment) or rear-view mirror (RVM) microphone (car environment). In each panel, the LER is shown as a function of window size T_w (in frames) for the sparse imputation (SI) method (solid line) with vertical bars around the data points indicating the 95% confidence intervals. The 95% confidence intervals for the cluster-based imputation (CI) method are indicated by dashed lines and that of the baseline recogniser (B) by dotted lines.

the lavalier microphone. Again, the car environment proves to be the more difficult recognition environment with a baseline LER score of 67.3 %, compared to 38.3 % in the public environment. As with the lavalier microphone data, the SI method performs substantially better than CI in the car environment, doing up to 11 % (absolute) better for window lengths in the range $T_w = [5, 15]$. In the public environment, SI also performs much better than CI, although by a smaller margin.

All in all, both CI and SI have a positive effect on recognition performance in comparison to the baseline, but clearly SI performs better in the more difficult conditions (lavalier and RVM microphone data from the car environment and the far-field microphone data from the public environments) at the cost of a small performance loss in relatively clean environments.

6.2. Artificially corrupted speech: babble noise scenario

Recognition performance using the clean speech data artificially corrupted with babble noise is displayed in Fig. 4. The baseline result for the clean speech signal is LER = 3.3 %.

In the SNR = 15 dB condition, SI and CI achieve comparable accuracies with LER $\approx$ 7.5 %, again with the exception of $T_w = 1$ when using an estimated missing data masks. We can also observe that using an error-free oracle masks leads to a much lower number of recognition errors: SI now comes closer to clean speech baseline recognition scores (LER = 4.1 %) while CI achieves 4.9 % LER. While in the case of the estimated mask, there is an indication of an optimum window length in the range $T_w = [10, 15]$, no such optimum can be seen when using the oracle mask.

In the SNR=10 dB condition, SI does significantly better than CI both for estimated and oracle missing data mask. The same can be observed in the SNR=5 dB condition, although the gap between SI and CI performance becomes much larger at lower SNR's, particularly when using the oracle mask. This indicates that SI gains more performance from the extra, error-free information contained in the oracle mask. In both SNR conditions, it is difficult to see a clear relation between the window length and SI performance, although LER seems to show a shallow minimum around $T_w = [10, 20]$. Remarkably, even in the $T_w = 1$ condition, SI does better than CI when using the oracle mask.

In the SNR=0 dB condition when using an estimated mask, neither SI nor CI can reconstruct the clean speech signal to a sufficient degree to achieve a usable performance , since both methods have LER around 75 %. When

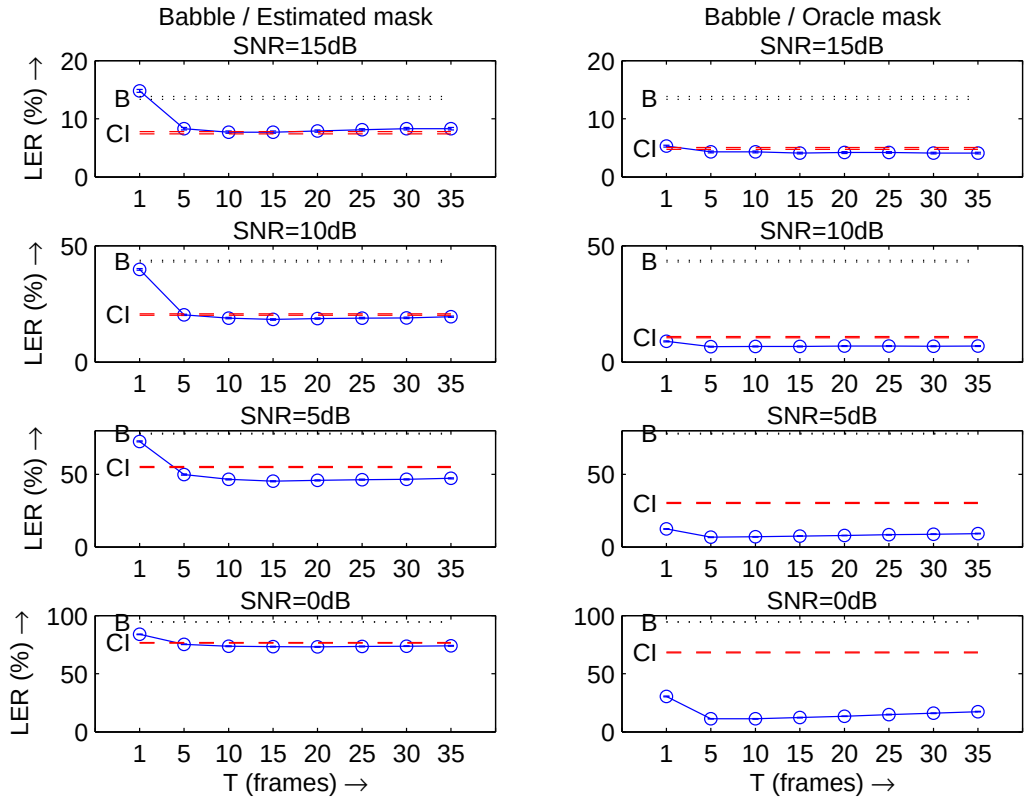


Figure 4: Recognition accuracy expressed as letter error rates (LER) for the dataset containing clean speech artificially corrupted with babble noise. The results are shown for the estimated missing data mask (left pane) and oracle missing data mask (right pane). Row one through four apply to different signal-to-noise ratio's (SNR). In each panel, the LER is shown as a function of window size T_w (in frames) for the sparse imputation (SI) method (solid line) with vertical bars around the data points indicating the 95% confidence intervals. The 95% confidence intervals for the cluster-based imputation (CI) method are indicated by dashed lines and that of the baseline recogniser (B) by dotted lines.

using an oracle mask, the situation is quite different. While CI achieves a LER of 68.3 %, gaining less than 10 % from using an oracle mask, SI achieves error rates of only 11.2 % at SNR = 0 dB.

7. Discussion

7.1. Sparse imputation for large vocabulary continuous speech recognition

Research on noise robust MDT started out with experiments on small vocabulary tasks artificially corrupted by noise [10, 5]. While vocabulary sizes have since increased, in the majority of cases still artificially corrupted speech has been employed [4, 47, 27].

It is not until recently that research has turned to MDT on large vocabulary speech recorded in realistic conditions [19, 48]. In this work, we investigated whether the improvements in recognition accuracy obtained with SI on the AURORA-2 digit recognition task [15] would generalise to a large vocabulary task.

Experiments on the SPEECON data indicate that SI is 1) indeed capable of significantly improving large vocabulary continuous speech recognition (LVCSR) performance on noisy speech data and 2) performs equally well on artificially corrupted data and noisy speech recorded in real-world environments where different types of microphones and different microphone–speaker distances were used.

Compared to the cluster-based imputation (CI) method, SI improved the speech recognition performance especially at low SNR's. When using an estimated missing data mask, SI performance on the cleanest conditions (headset-recorded data or artificially corrupted data at SNR=15 dB) was comparable or slightly lower than the CI performance, but SI performed better than CI in all the noisier conditions. The difference between SI and CI was most notable when using SI on the car data recorded with a rear-view-mirror (RVM) microphone, which resulted in 26 % relative error reduction in LER. In the experiments using oracle masks, SI outperformed CI at all SNR levels, and as with the estimated masks, the differences in performance grew larger at low SNR's. On the artificially corrupted data at SNR=0 dB, the relative error reduction in LER was 84 % compared to the CI performance.

7.1.1. Sparse representation of speech

Initially, we expressed some concern regarding the performance limits of an exemplar-based method in modelling large vocabulary speech, especially

when using windows spanning multiple time frames. The oracle mask results reported in Section 6, however, serve as empirical evidence that such concerns are unfounded, since the use of 8000 exemplars suffices for SI to find a proper reconstruction of the underlying clean speech, even when many features are missing.

To study this issue in more detail, we did a small additional experiment in which we investigated for three window lengths the sparsity of clean (uncorrupted) speech of a random subset of 10 utterances of the SPEECON test database. At window lengths $T_w \in \{10, 20, 30\}$, the utterances contain $6794, 6694, 6594$ windows, respectively.

For each of the windows, the observation vector was first normalised to a Euclidean unit norm after which we recovered its sparse representation $\boldsymbol{x}$ using the `l1_ls_nonneg` solver solver (cf. Section 5.6) and sorted the elements of this vector with respect to weight. Finally, we averaged the sorted weight vectors over all windows.

The result is an average weight vector ordered with respect to weight which indicates how many exemplars (on average) are needed to represent the windows of the selected test utterances. For each window length, the 45 largest weights are shown in Figure 5. From this figure it can be deduced that the fixed length spectrogram segments of large vocabulary SPEECON speech in the test set can indeed be sparsely represented. The results show that, within the accuracy of the solver, on average windowed spectra can be sparsely represented using no more than approximately 30 exemplars.

7.1.2. Low-noise conditions

With CI performance being consistently lower than the SI performance in all the noisier conditions, it is interesting to explore in more detail why CI performs comparable or better than SI when using an estimated mask in the cleanest conditions, i.e., the headset-recorded data or artificially corrupted data at SNR=15 dB. A factor that may contribute to the difference is that CI uses the unreliable features as an upper bound during the missing feature reconstruction. This has been shown to improve the MDT performance in various noise conditions [10]. In SI, the upper bound is applied on the reconstructed features *after* selection of the exemplars as indicated in Equation (10). As a consequence, it is not taken into account that if some of the estimates are considered incorrect because they are larger than the observed upper bound, there is no guarantee the other features are correct since they stem from the same linear combination of exemplars.

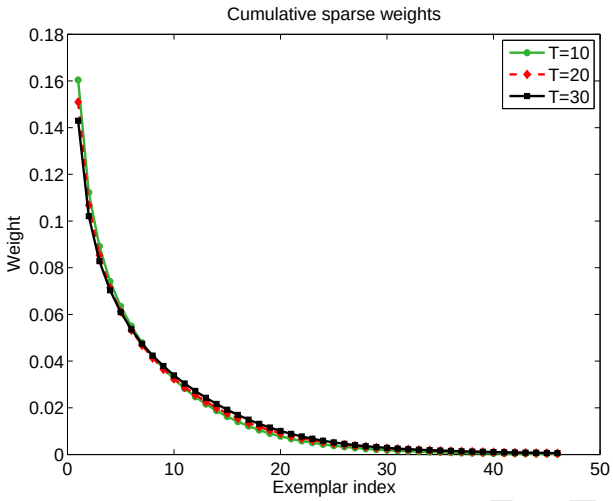


Figure 5: The sparsity of clean speech in a subset of the SPEECON test database. The graph shows the average weight of the 45 largest nonzero elements of $\boldsymbol{x}$ of each sparsely represented window in a random subset of 10 utterances. The results displayed here pertain to the window lengths $T_w \in \{10, 20, 30\}$.

The upper bounds are likely to have the largest impact in low-noise conditions, which is where CI performs best, but the treatment of upper bounds does not explain why SI does better than CI at SNR=15 dB if the estimated masks are replaced with oracle masks. Analysing the recognition errors in detail (not shown) revealed that when CI outperforms SI, the difference is mainly due to insertion errors. This is in line with the observations in [15]. The SI method is prone to insertion errors because even a single, isolated, reliable feature results in the entire T_w-frame window being represented as a linear combination of clean speech exemplars. Depending on the decoder, such a single reliable feature could lead to the insertion of a segment/letter/word. These insertion errors are most likely when using an estimated mask that erroneously marks spurious features reliable.

It is noteworthy that in the noisier conditions, an isolated reliable feature may be all that is left of the underlying clean speech, and indeed, the same property that makes SI prone to insertion errors in relatively clean conditions is what causes the performance gain in the more difficult noise conditions where SI performs better than CI. A solution for this issue may

be found in techniques such as weighted Viterbi decoding [13] or uncertainty decoding [49], in which the decoder takes the uncertainty of the accuracy with which clean speech was estimated into account. Preliminary experiments on artificially corrupted speech showed that SI performance indeed increases both at high and low SNR's when using a proper measure of uncertainty (cf. [50]).

7.2. Optimal time context

7.2.1. Using multiple frames of time-context

Figures 3 and 4 in Section 6 clearly indicate that increasing the time context beyond $T_w = 1$ in the SI method improves the speech recognition performance. In most conditions, the optimal recognition performance is achieved when the time context in SI is approximately 5–20 frames. A detailed analysis of the recognition errors (not shown) revealed that increasing the time context past the optimum range systematically leads to an increased number of deletion errors and that in general, when SI outperformed CI, it was due to a reduced number of deletion errors.

Analysis of the clean estimates produced by SI at window sizes larger than 5–20 frames (not shown) revealed that the increased number of deletions is due to difficulties in finding a sparse linear combination of exemplars that would describe the high-dimensional observation sufficiently accurately. Consequently, the resulting sparse representations only capture the high energy regions of the spectrogram window and do not describe the details accurately. In addition, the number of spectrograms over which we average to calculate the final estimate increases with increasing window length, results in smoothed approximations for the clean speech spectrograms and tends to decrease the dynamic range of the feature values. Since the decoder employs mean normalisation, the smoothed, mean-normalised features result in a reduced contrast between states and may even start to resemble silence. This, in turn, can lead to recognition errors such as deletions.

Finally, as the results reported in Section 6 and the previous results in [14] indicate that missing data imputation using SI benefits from increasing the time context, the question could come to mind whether using more time-context might also be advantageous for CI. While we do not address this question in this work, it should be noted that the computational complexity of CI is $\mathcal{O}(K^4)$, where K denotes the feature dimension, which shows that adding time context (effectively increasing the number of features per frame) would quickly become infeasible. Moreover, there is a quadratic increase

in the amount of data required to accurately estimate the full covariances needed for CI, which can lead to training data scarcity.

7.2.2. SI performance for single time frames

Although the SI performance is not better than CI performance for every noise scenario and window length T_w, the performance is never much worse than CI, except for $T_w = 1$ in combination with an estimated mask. In this case, the SI performance is occasionally even worse than the uncompensated baseline system performance. However, if oracle masks are used instead, the SI performance is substantially better than CI, even with $T_w = 1$.

There are two issues that could explain the difference. First, the way frames are treated when all features are unreliable is different in SI and CI. In SI, frames without any reliable features are imputed as silence. If several of such frames occur during a speech segment, the decoder is more or less forced to recognise these segments as silence. Since it becomes less likely that a speech segment of T_w frames does not contain any reliable frames when T_w is large, SI achieves better results at longer window sizes. In contrast, CI imputes frames without any reliable values by making use of the prior probability given the noisy data. In this approach, the imputed frame is less likely to be interpreted as silence all the time.

To test this hypothesis, we ran an additional experiment on the artificial babble noise data. Here, CI was modified to also impute silence for all the frames that contain no reliable features. The results (not shown) confirmed that this decreases the system performance when an estimated missing data mask is used, although the performance was still better than the SI performance at $T_w = 1$. Interestingly, the modification *improved* the CI performance when an oracle mask was used, although the obtained accuracies were still substantially lower than the SI accuracies at $T_w = 1$. In summary, this experiment shows not only that imputing silence when all features are unreliable is only a good approach in the absence of mask estimation errors, but also that the difference in treating frames without any reliable values does not fully explain the differences between SI using $T_w = 1$ and CI.

A second factor that may contribute to the sub-optimal performance of SI when $T_w = 1$ and an estimated mask is used, is that at $T_w = 1$, the minimisation problem (9) gets extremely under-determined as the number of reliable features decreases. It is a property of the applied technique that the quality of the reconstruction does not deteriorate gradually, but suffers from a sudden break-down once the number of reliable features gets below

a certain threshold. This happens because with very few reliable values, the sparsest representation can suddenly be constituted of rather arbitrary exemplar vectors. As discussed in detail in [14], it is very difficult to give an estimate on the minimum number of (reliable) features needed for imputation, but the results in the paper can be seen as an empirical indication that in low SNR conditions, a sizable proportion of single time frames does not contain enough reliable features for the SI method to work properly.

In light of these results, we may need to reconsider the SI approach to impute silence for frames that contain no reliable features, at least when a short window is used. Possibly, a better approach would be to interpolate between neighbouring frames, to use the output of the CI method for frames without any reliable values, or to use an approach related to the uncertainty decoding approach mentioned above by setting all state likelihoods to an equal value for frames which do not contain reliable features.

7.3. Influence of mask quality

It is well known that recognition accuracy improvements obtainable with MDT are highly dependent on the quality of the applied missing data mask. When comparing the CI results for estimated and oracle mask, it is obvious that the absence of mask estimation errors substantially improves recognition accuracy, with relative LER improvements ranging from 10% to 45%. The same holds for sparse imputation, although the differences are larger, particularly at low SNR's. As can be seen in Figure 4, at an SNR of 0 dB, SI achieves an accuracy of 11.2 % LER as opposed to the 73.1% with the estimated mask at $T_w = 20$. As in our previous experiments on a digit recognition task [15], the large performance gap between the two types of missing data masks indicates that SI can potentially perform much better than CI, provided mask estimation errors are reduced.

An equally valid conclusion is that SI is more sensitive to mask estimation errors than CI. The reason for this is that mask estimation errors which incorrectly label features reliable will mislead the search for the 'true' sparse representation associated with the underlying clean speech. Depending on the location of these features, imputed features can become very different from the underlying clean speech. Using a more conservative estimation method is no solution, since mask estimation errors which incorrectly label features as unreliable reduce the overall number of reliable features available for imputation and cause the search to miss out on features useful for distinguishing between exemplars [14].

The recognition performance using error-free oracle masks suggest an even larger real-world potential for SI, provided substantially better mask estimation methods can be found. Alternatively, the SI method could be made more robust against mask estimation errors by a number of algorithmical improvements. An attractive approach would be to change the search for exemplars, i.e., the minimisation problem (9), to include the constraint that the reconstructed speech should not exceed the noisy observation. Other possibilities for improvements are the additional use of derivative features in the exemplars rather using only static features, or the use of *soft* missing data masks [51]. In soft masks the binary reliability score is replaced by the probability that a spectral component is reliable, providing more robustness against mask estimation errors. For a more extensive discussion on these potential improvements, we refer the reader to [14, 52].

7.4. Computational effort

To roughly characterise the computational effort needed, we did a small test of the running time of the CI algorithm on a machine with a Core 2 Duo E6550 2.33 GHz processor. The running time for an utterance of 756 frames (6 seconds of speech) containing 11 734 unreliable values was 80 seconds. For the SI algorithm with a window length of $T_w = 10$ frames, the running time on this utterance was 61 seconds.

While SI is faster for this particular utterance, SI still performs at about 10 times real-time. It is therefore interesting to study its computational complexity. SI using the `SolveLasso` solver has a computational complexity of $\mathcal{O}\left(W\left((RT_w)^3 + NRT_w\right)\right)$, where W denotes the number of windows to be processed, N the number of clean speech exemplars in the dictionary, R the average number of reliable features per frame, and T_w the window length in frames. In practice, the computational complexity is completely dominated by the term NRT_w.

There are three ways to reduce the computational effort, each of which scale approximately linearly. The first is to increase the window shift Δ which decreases the number of windows W. In [15] it was shown that increasing Δ too much reduces imputation performance, but also that small increases do not decrease accuracy. Moreover, Δ does not have to be constant over the utterance and could for example be made dependent on the number of reliable features. The second approach is to reduce the number of features in each window. Aside from reducing the window length, which in Section 7.2 was shown to decrease performance when reduced to much, it is also possible

to apply dimensionality reduction to the features in a window, as used in [33]. Finally, the dictionary size might be reduced by methods such as clustering and an algorithmic way to handle shift-invariance rather than by including time-shifted variants of the same phenomena.

8. Conclusions and future work

In this work, we investigated the performance of the sparse imputation missing data technique on read sentences of the Finnish SPEECON corpus using real-world noisy speech recordings as well as clean speech recordings artificially corrupted with babble noise. In previous research sparse imputation was shown to be an effective method for improving the noise robustness of a connected digit recognizer using the artificially noisifid AURORA-2 data [15] . The current results show that the method can be readily extended to a large vocabulary recognition task in which the speech suffers from corruptions found in real-world environments. We found that also in the SPEECON corpus, fixed-length spectrogram windows can be adequately represented by a sparse, linear combination of exemplars: On average less than 30 are needed. As with the AURORA-2 task, sparse imputation on SPEECON greatly benefits from using additional time-context for imputation. With a dictionary size of 8000 randomly selected exemplars used in this study, we typically found a context of 5-20 frames (i.e., 50-200 ms) to yield the best recognition accuracies.

Experiments on artificially corrupted speech indicated that sparse imputation outperforms a conventional imputation technique by a significant margin when the ideal 'oracle' reliability of noisy speech features are used. With error-prone reliability estimates, sparse imputation performs slightly worse than our baseline imputation technique in the cleanest conditions, but significantly better at lower SNR's.

Future work in the sparse imputation framework will focus on improving the robustness toward mask estimation errors and better handling the frames which do not contain any reliable values.

Acknowledgement

The research by Jort Florent Gemmeke was carried out in the MIDAS project, granted under the Dutch-Flemish STEVIN program. The research by Ulpu Remes was supported by the Helsinki Graduate School in Computer

Science and Engineering and by the Academy of Finland in the projects *Auditory approaches to automatic speech recognition* and *Adaptive Informatics Research Centre*. We acknowledge Lou Boves and Kalle Palomäki for useful discussions.

References

[1] B. Raj, R. M. Stern, Missing-feature approaches in speech recognition, IEEE Signal Processing Magazine 22 (5) (2005) 101–116.

[2] M. Cooke, P. Green, M. Crawford, Handling missing data in speech recognition, in: Proc. ICSLP, Yokohama, Japan, 1994, pp. 1555–1558.

[3] B. Raj, R. Singh, R. Stern, Inference of missing spectrographic features for robust automatic speech recognition, in: Proc. ICSLP, Sydney, Australia, 1998, pp. 1491–1494.

[4] B. Raj, M. Seltzer, R. Stern, Reconstruction of missing features for robust speech recognition, Speech Communication 43 (4) (2004) 275–296.

[5] H. Van hamme, PROSPECT features and their application to missing data techniques for robust speech recognition, in: Proc. INTERSPEECH, Jeju Island, Korea, 2004, pp. 101–104.

[6] L. Josifovski, M. Cooke, P. Green, A. Vizinho, State based imputation of missing data for robust speech recognition and speech enhancement, in: Proc. EUROSPEECH, Budapest, Hungary, 1999, pp. 2837–2840.

[7] M. V. Segbroeck, Robust large vocabulary continuous speech recognition using missing data techniques, Ph.D. thesis, K.U.Leuven (2010).

[8] H. Van hamme, Robust speech recognition using cepstral domain missing data techniques and noisy masks, in: Proc. ICASSP, Montreal, Quebec, Canada, 2004, pp. 213–216.

[9] V. Stouten, Robust automatic speech recognition in time-varying environments, Ph.D. thesis, K.U.Leuven (2006).

[10] M. Cooke, P. Green, L. Josifovksi, A. Vizinho, Robust automatic speech recognition with missing and unreliable acoustic data, Speech Communication 34 (3) (2001) 267–285.

[11] J. Gemmeke, B. Cranen, Using sparse representations for missing data imputation in noise robust speech recognition, in: Proc. EUSIPCO, Lausanne, Switzerland, 2008.

[12] B. Raj, Reconstruction of incomplete spectrograms for robust speech recognition, Ph.D. thesis, Carnegie Mellon University (2000).

[13] Z.-H. Tan, P. Dalsgaard, B. Lindberg, Exploiting temporal correlation of speech for error-robust and bandwidth-flexible distributed speech recognition, IEEE Transactions on Audio, Speech and Language Processing 15 (4) (2007) 1391–1403.

[14] J. F. Gemmeke, H. V. hamme, B. Cranen, L. Boves, Compressive sensing for missing data imputation in noise robust speech recognition, IEEE Journal of Selected Topics in Signal Processing 4 (2) (2010) 272–287.

[15] J. Gemmeke, B. Cranen, Missing data imputation using compressive sensing techniques for connected digit recognition, in: Proc. DSP, Santorini, Greece, 2009, pp. 1–8.

[16] H. Hirsch, 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, 2000, pp. 181–188.

[17] D. Iskra, B. Grosskopf, K. Marasek, H. V. D. Heuvel, F. Diehl, A. Kiessling, SPEECON - speech databases for consumer devices: Database specification and validation, in: Proc. LREC, Las Palmas, Canary Islands, Spain, 2002, pp. 329–333.

[18] A. Varga, H. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems, Speech Communication 12 (3) (1993) 247–51.

[19] U. Remes, K. J. Palomäki, M. Kurimo, Missing feature reconstruction and acoustic model adaptation combined for large vocabulary continuous speech recognition, in: Proc. EUSIPCO, Lausanne, Switzerland, 2008.

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

[21] J. Barker, M. Cooke, D. Ellis, Decoding speech in the presence of other sources, Speech Communication 45 (1) (2005) 5–25.

[22] A. Nadas, D. Nahamoo, M. Picheny, Speech recognition using noise-adaptive prototypes, IEEE Transactions on Acoustics, Speech and Signal Processing 37 (10) (1989) 1495–1503.

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

[24] M. Seltzer, B. Raj, R. Stern, A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition, Speech Communication 43 (4) (2004) 379–393.

[25] W. Kim, R. M. Stern, Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise, in: Proc. ICASSP, Toulouse, France, 2006, pp. 305–308.

[26] R. Weiss, D. Ellis, Estimating single-channel source separation masks: Relevance Vector Machine classifiers vs. pitch-based masking, in: Proc. Workshop on Statistical and Perceptual Audition SAPA-06, Pittsburgh, Pennsylvania, USA, 2006, pp. 31–36.

[27] M. Van Segbroeck, H. Van hamme, Vector-Quantization based mask estimation for missing data automatic speech recognition, in: Proc. INTERSPEECH, Antwerp, Belgium, 2007, pp. 910–913.

[28] S. Harding, J. Barker, G. J. Brown, Mask estimation for missing data speech recognition based on statistics of binaural interaction, IEEE Transactions on Audio, Speech and Language Processing 14 (1) (2006) 58–67.

[29] N. Ma, P. Green, J. Barker, A. Coy, Exploiting correlogram structure for robust speech recognition with multiple speech sources, Speech Communication 49 (12) (2007) 874–891.

[30] C. Cerisara, S. Demange, J.-P. Haton, On noise masking for automatic missing data speech recognition: A survey and discussion, Computer Speech and Language 21 (3) (2007) 443–457.

[31] J. Barker, M. Cooke, 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, 2001, pp. 213–216.

[32] S. B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (4) (1980) 357–366.

[33] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2009) 210–227.

[34] D. L. Donoho, Compressed sensing, IEEE Transactions on Information Theory 52 (4) (2006) 1289–1306.

[35] 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 59 (6) (2006) 797–829.

[36] E. J. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Communications On Pure and Applied Mathematics 59 (8) (2006) 1207–1223.

[37] J. Pylkkönen, An efficient one-pass decoder for Finnish large vocabulary continuous speech recognition, in: Proc. 2nd Baltic Conference on Human Language Technologies, Tallinn, Estonia, 2005, pp. 167–172.

[38] J. J. Odell, The use of context in large vocabulary speech recognition, Ph.D. thesis, University of Cambridge (1995).

[39] J. Pylkkönen, M. Kurimo, Duration modeling techniques for continuous speech recognition, in: Proc. INTERSPEECH, Jeju Island, Korea, 2004, pp. 385–388.

[40] M. Creutz, K. Lagus, Unsupervised discovery of morphemes, in: Proc. ACL-02 Workshop on Morphological and Phonological Learning, Philadelphia, Pennsylvania, USA, 2002, pp. 21–30.

[41] CSC Tieteellinen laskenta Oy, The language bank of Finland (2001). URL www.csc.fi/languagebank/

[42] V. Siivola, B. Pellom, Growing an n-gram language model, in: Proc. INTERSPEECH, Lisbon, Portugal, 2005, pp. 1309–1312.

[43] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, J. Pylkkönen, Unlimited vocabulary speech recognition with morph language models applied to Finnish, Computer Speech and Language 20 (4) (2006) 515–541.

[44] M. Gales, Semi-tied covariance matrices for hidden Markov models, IEEE Transactions on Speech and Audio Processing 7 (3) (1999) 272–281.

[45] M. Cooke, A glimpsing model of speech perception in noise, J. Acoust. Soc. Am. 119 (3) (2006) 1562–1573.

[46] S. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, An interior-point method for large-scale l1-regularized least squares, IEEE Journal on Selected Topics in Signal Processing 1 (4) (2007) 606–617.

[47] S. Srinivasan, N. Roman, D. Wang, Binary and ratio time-frequency masks for robust speech recognition, Speech Communication 48 (11) (2006) 1486–1501.

[48] J. F. Gemmeke, Y. Wang, M. V. Segbroeck, B. Cranen, H. V. hamme, Application of noise robust MDT speech recognition on the SPEECON and SpeechDat-Car databases, in: Proc. INTERSPEECH, Brighton, UK, 2009.

[49] H. Liao, M. J. F. Gales, Issues with uncertainty decoding for noise robust automatic speech recognition, Speech Communication 50 (4) (2008) 265–277.

[50] J. F. Gemmeke, U. Remes, K. J. Palomäki, Observation uncertainty measures for sparse imputation, in: Submitted to INTERSPEECH 2010.

[51] J. Barker, L. Josifovski, M. Cooke, P. Green, Soft decisions in missing data techniques for robust automatic speech recognition, in: Proc. ICSLP, Beijing, China, 2000, pp. 373–376.

[52] J. Gemmeke, B. Cranen, Sparse imputation for noise robust speech recognition using soft masks, in: Proc. ICASSP, Taipei, Taiwan, 2009, pp. 4645–4648.