Exemplar-based sparse representations for noise robust automatic speech recognition

Jort F. Gemmeke*, *Student-Member, IEEE,* Tuomas Virtanen, Antti Hurmalainen

***Abstract*—This paper proposes to use exemplar-based sparse representations for noise robust automatic speech recognition. First, we describe how speech can be modelled as a linear combination of a small number of exemplars from a large speech exemplar dictionary. The exemplars are time-frequency patches of real speech, each spanning multiple time frames. We then propose to model speech corrupted by additive noise as a linear combination of noise and speech exemplars, and we derive an algorithm for recovering this sparse linear combination of exemplars from the observed noisy speech. We describe how the framework can be used for doing hybrid exemplar-based/HMM recognition by using the exemplar-activations together with the phonetic information associated with the exemplars.**

As an alternative to hybrid recognition, the framework also allows us to take a source separation approach which enables exemplar-based feature enhancement as well as missing data mask estimation. We evaluate the performance of these exemplar-based methods in connected digit recognition on the AURORA-2 database. Our results show that the hybrid system performed substantially better than source separation or missing data mask estimation at lower SNRs, achieving up to 57.1% accuracy at SNR= -5 dB. Although not as effective as two baseline recognisers at higher SNRs, the novel approach offers a promising direction of future research on exemplar-based ASR.

***Index Terms*—Speech recognition, exemplar-based, noise robustness, sparse representations, non-negative matrix factorisation**

I. Introduction

FOR the last 30 years Automatic Speech Recognition (ASR) has been dominated by the use of Hidden Markov Models (HMMs) employing Gaussian Mixture Models (GMMs) to model the statistics of the acoustics [1]. The ASR performance of these systems, however, degrades substantially when speech is corrupted by background noise not seen during training. The reason for this is that the observed speech signal does no longer match the distributions derived from the training material.

There have been numerous approaches that aim at resolving this mismatch, such as normalisation or enhancement of the speech features [1], compensation of the acoustic models [2], [3] and the use of recogniser architectures that use only the least noisy observations [4], to name a few. Originally, most noise robustness techniques were based on strong stationarity assumptions about the underlying noise, but methods have been proposed that address non-stationary noise [5]–[8].

Recently, models based on *sparse representations* have gained considerable interest in signal processing. Sparse representations are representations that account for most or all information of a signal with a linear combination of only a small number of elementary signals, called atoms. The collection of atoms that is used is called a *dictionary*. In audio, sparse representations have been used for source separation by expressing a signal that is a mixture of multiple sources with a sparse representation, using a dictionary for each underlying source. That sparse representation is determined by finding the sparsest possible linear combination that describes the observed signal, using techniques best known from the fields of non-negative matrix factorisation (NMF) [9] and Compressed Sensing [10]. Reconstruction using parts of the dictionary pertaining to only a single source, results in an estimate of the underlying source [11]–[15].

Another application of sparse representations is pattern recognition. This is done by associating the dictionary atoms with class labels, and using the weights of the atoms in the sparse representation as evidence for the class of the observed signal. This approach has lead to state-of-the-art classification results in various fields, such as face recognition [16] and phone classification [17].

In this work, we investigate the effectiveness of combining these two approaches. Expressing noisy speech as a sparse linear combination of speech and noise dictionary atoms, we first determine the sparse representation. With the atoms in the speech part of the dictionary associated with speech labels, we can then use the weights of the speech part of the sparse representation to provide noise robust evidence for the identity of the underlying speech unit. Based on preliminary experiments in [18], [19], we propose to use this approach, dubbed *sparse classification* (SC) in earlier work [20], in a hybrid SC/HMM speech recogniser. Hybrid HMM systems are commonly used when replacing GMM-based modelling of the acoustics by alternative modelling techniques, such as neural network based systems [1].

In most speech applications of sparse representations, the dictionary atoms either consist of fundamental basis functions such as Fourier coefficients or wavelets, or are *learned* [11]–[14], [21]. In this work, however, we model signals as a sparse linear combinations of *examples* of that signal [16]. Thus, we model speech segments as a weighted linear combination of

The research by Jort F. Gemmeke was carried out in the MIDAS project. The MIDAS project is carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (http://www.stevin-tst.org). The research of Tuomas Virtanen and Antti Hurmalainen has been funded by the Academy of Finland.

Jort F. Gemmeke is with the Centre for Language and Speech Technology, Radboud University Nijmegen, NL-6500 HD Nijmegen, The Netherlands, e-mail: J.Gemmeke@let.ru.nl

Tuomas Virtanen and Antti Hurmalainen are with the Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland, e-mail: tuomas.virtanen@tut.fi, antti.hurmalainen@tut.fi

Manuscript received ?; revised ?

example speech segments, *exemplars* [15], [18], [22]. These exemplars are spectrographic representations of speech spanning multiple time-frames of speech (typically 50 to 300 ms). The use of exemplars to model speech is reminiscent of more traditional exemplar-based approaches to speech recognition [1], [23]. In those approaches, however, speech is represented by one or more exemplars that each individually have the smallest distance to the observed speech token, whereas in our framework, speech exemplars *jointly* approximate the observed speech.

The use of a speech dictionary containing exemplars as atoms has several advantages. First, the dictionary is relatively easy to construct by extraction of speech segments from a speech database. Second, it becomes computationally efficient to construct dictionaries with high-dimensional atoms that contain several frames of time context, which makes confusion between noise and speech atoms in less likely. Third, the dictionary can allow for very sparse representations if an observed speech segment closely resembles speech contained in the dictionary [22]. Finally, the use of exemplars makes the mapping from atoms to speech classes straightforward: Each time-frame in the speech exemplars is directly labelled with an HMM-state label, obtained by means of a forced alignment of the transcription on the training database using a conventional HMM-based recogniser.

In the SC approach, the weights of the linear combination of speech exemplars are used to provide a weighted sum of HMM-state scores for each frame in the observed speech. In order to investigate the effectiveness of the SC approach, we also use the exemplar-based sparse representations to apply two conventional robust ASR techniques. First, we extend the sparse representation-based source separation approach described earlier to allow the use of atoms that span multiple time-frames. We can then do feature enhancement, which aims at providing clean speech features which are recognised by a conventional, GMM/HMM-based ASR system.

Second, we use the exemplar-based sparse representations to apply a missing data technique (MDT) [4], [5]. In noisy speech, MDT distinguishes between features dominated by speech ('reliable' features) and features dominated by noise ('unreliable' c.q. 'missing'). Discarding the unreliable features, speech recognition (using acoustic models trained on clean speech) is done on the incomplete data by imputation or marginalisation of the missing features. We create this reliable/unreliable labelling of noisy speech features by comparing the speech and noise estimates provided by our source separation result [24].

The main contributions of this work are twofold. First, we investigate the effectiveness of combining two techniques employing sparse representations, source separation and classification, and second, we investigate to what extent using dictionary atoms that span multiple frames is beneficial for sparse representation-based noise robustness techniques. We compare the recognition accuracies of the various approaches using material from the AURORA-2 database, which contains connected digits artificially corrupted by a number of different noises at several SNRs. In order to investigate the influence of using exemplars spanning multiple time-frames, we investigate recognition accuracy as a function of exemplar size. We compare the performance of the three exemplar-based approaches to a multi-condition trained recogniser and to a noise-robust MDT approach in which the reliability estimates are based on a harmonic decomposition of speech [25].

II. Model for noisy speech

A. Sparse representation of noisy speech

Speech signals are represented by their spectro-temporal distribution of acoustic energy, a *spectrogram*. The exemplar-based approaches proposed in this paper operate in the Mel-scale magnitude spectrogram domain, with the term *magnitude* referring to the square root of energy in a time-frequency element. The cepstral features used in conventional ASR systems are based on a (mostly logarithmic) compression of the magnitude values followed by a decorrelating cosine transform. In our framework, however, we use the magnitude values directly to simplify the additivity of speech and noise.

The magnitude spectrogram describing a clean speech signal is a $B \times T$ dimensional matrix $\mathbf{S}$ (with B frequency bands and T time frames). To simplify the notation, the columns of this matrix are stacked into a single vector $\mathbf{s}$ of length $E = B \cdot T$, so that the entry $S(b,t)$, with $1 \leq b \leq B$ and $1 \leq t \leq T$, corresponds to the entry $s(b + (t-1)B)$.

We assume that an arbitrary speech spectrogram $\mathbf{s}$ can be expressed as a linear, non-negative combination of clean speech exemplars $\mathbf{a}^{\mathrm{s}}_j$, with $j = 1, \ldots, J$ denoting the exemplar index. These exemplars are magnitude spectrograms describing segments of speech signals extracted from a training database and are stacked in same way as was done to obtain $\mathbf{s}$. We write:

$$\mathbf{s} \approx \sum_{j=1}^{J} \mathbf{a}^{\mathrm{s}}_j x^s_j = \mathbf{A}^{\mathrm{s}} \mathbf{x}^{\mathrm{s}} \quad \text{subject to} \quad \mathbf{x}^{\mathrm{s}} \geq 0 \tag{1}$$

with x^s_j being the non-negative weight or *activation* of each exemplar. In this paper, the superscript s denotes speech, and the superscript n will denote noise. The J exemplars $\mathbf{a}^{\mathrm{s}}_1$, $\mathbf{a}^{\mathrm{s}}_2, \ldots, \mathbf{a}^{\mathrm{s}}_J$ are grouped into a speech exemplar matrix $\mathbf{A}^{\mathrm{s}}$ as $\mathbf{A}^{\mathrm{s}} = [\mathbf{a}^{\mathrm{s}}_1 \ \mathbf{a}^{\mathrm{s}}_2 \ldots \mathbf{a}^{\mathrm{s}}_J]$ and the activations stacked into $\mathbf{x}^{\mathrm{s}}$, a J-dimensional activation vector.

Previous research has shown that $\mathbf{x}^{\mathrm{s}}$ can be extremely *sparse* [22]. That is, only a few non-zero entries suffice to represent $\mathbf{s}$ with sufficient accuracy. The activations are restricted to non-negative values, a restriction which has turned out to be critical in audio analysis algorithms employing magnitude spectrograms [12].

Like clean speech, we assume we can model a $B \times T$ dimensional noise spectrogram $\mathbf{N}$, represented by the stacked vector $\mathbf{n}$, as a linear combination of K noise exemplars $\mathbf{a}^{\mathrm{n}}_k$, with $k = 1, \ldots, K$ being the noise exemplar index. We can now represent a noisy speech segment $\mathbf{Y}$, reshaped into vector $\mathbf{y}$, as a linear combination of both speech and noise exemplars:

$$\mathbf{y} \approx \mathbf{s} + \mathbf{n} \tag{2}$$

$$\approx \sum_{j=1}^{J} \mathbf{a}_j^{\mathrm{s}} x_j^s + \sum_{k=1}^{K} \mathbf{a}_k^{\mathrm{n}} x_k^{\mathrm{n}} \tag{3}$$

$$= [\mathbf{A}^{\mathrm{s}} \mathbf{A}^{\mathrm{n}}] \begin{bmatrix} \mathbf{x}^{\mathrm{s}} \\ \mathbf{x}^{\mathrm{n}} \end{bmatrix} \quad \text{s.t.} \quad \mathbf{x}^{\mathrm{s}}, \mathbf{x}^{\mathrm{n}} \geq 0 \tag{4}$$

$$= \mathbf{A}\mathbf{x} \quad \text{s.t.} \quad \mathbf{x} \geq 0 \tag{5}$$

where $\mathbf{x}^{\mathrm{n}}$ is the activation vector of the noise exemplars and $\mathbf{A}^{\mathrm{n}}$ is the matrix containing noise exemplars. The whole speech + noise exemplar dictionary matrix $\mathbf{A}$ has dimensionality $E \times L$, where $L = J + K$, and vector $\mathbf{x}$ contains the activations of the speech and noise exemplars. Since $\mathbf{x}$ is assumed to be sparse, and represents the noisy observation in terms of the exemplar activations, $\mathbf{x}$ is referred to as a *sparse representation*.

We normalise the dictionary rows so each frequency band has the same weight, and normalise the dictionary columns (corresponding to exemplars) because this has been found to produce slightly better results in source separation [14]. Normalisation is done by iteratively scaling each row and column so that its Euclidean norm equals unity. After normalisation, the norms of the columns equal unity, and the norms of the rows are approximately equal. During decoding, each noisy speech segment $\mathbf{y}$ is scaled using the frequency band normalisation applied to $\mathbf{A}$. Because the magnitude of the exemplar activations in $\mathbf{x}$ can vary, arbitrary speech levels and SNRs can be matched.

B. Finding the activations

In order to obtain the sparse representation $\mathbf{x}$, we search for linear combinations of exemplars which are able to represent the noisy speech $\mathbf{y}$ with the model $\mathbf{A}\mathbf{x}$, while using only a small number of nonzero entries in $\mathbf{x}$. We give a visual example of this process in Fig. 1. The linear combination of exemplars is found by minimising the cost function:

$$d(\mathbf{y}, \mathbf{A}\mathbf{x}) + ||\boldsymbol{\lambda}. * \mathbf{x}||_p \quad \text{s.t.,} \quad \mathbf{x} \geq 0 \tag{6}$$

The first term measures the distance between the noisy observation and the model using function d. The second term enforces sparsity by penalising the non-zero entries of $\mathbf{x}$ using the L_p norm of the activation vector, weighted by element-wise multiplication (operator $.*$) of the vector $\boldsymbol{\lambda} = [\lambda_1 \quad \lambda_2 \ldots \lambda_L]$. The activations of the exemplars are constrained to be non-negative.

We use the generalised Kullback-Leibler (KL) divergence for d:

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{e=1}^{E} y_e \log(\frac{y_e}{\hat{y}_e}) - y_e + \hat{y}_e \tag{7}$$

In source separation methods, the KL divergence has been found to produce better results than for example the Euclidean distance [12]. The statistical interpretation behind the minimisation of (6) can be found in [26].

We choose to control the sparseness used in the second term of (6) by the L_1 norm, which has been found to be effective in obtaining sparse solutions (cf. [10] and the references therein): $||\boldsymbol{\lambda}.*\mathbf{x}||_1 = \sum_{l=1}^{L} x_l \lambda_l$. Unlike in most studies, where a single scalar weight is used to penalise all non-zero entries equally, we allow different weights for speech and noise exemplars in the dictionary. In pilot experiments, it was found that enforcing the sparseness of speech exemplars was very important. The reason for this is that the linear combination of exemplars is naturally sparse because an observed speech segment should ideally be represented only by exemplars pertaining to the same underlying speech unit [16]. Therefore, enforcing the sparsity of the speech exemplars results (albeit indirectly) in activation of exemplars which represent the same underlying speech unit as the observed speech segment. We do not make such assumptions about the corrupting noise and thus do not enforce the sparseness of the noise exemplars.

The cost function (6) is minimised by first initialising the entries of the vector $\mathbf{x}$ to unity, and then iteratively applying the update rule:

$$\mathbf{x} \leftarrow \mathbf{x}. * (\mathbf{A}^{\mathrm{T}}(\mathbf{y}./(\mathbf{A}\mathbf{x})))./(\mathbf{A}^{\mathrm{T}}\mathbf{1} + \boldsymbol{\lambda}). \tag{8}$$

with $.*$ and $./$ denoting element-wise multiplication and division, respectively. The vector $\mathbf{1}$ is an all-one vector of length E. The derivation of (8) is given in appendix A.

C. Sliding window approach for time-continuity

Describing speech as a linear combination of exemplars is only feasible for relatively short signal segments. In this work, we consider segments with a duration of 50 to 300 ms. In order to decode utterances of arbitrary lengths, we adopt a sliding time window approach as in [18]. We first divide an utterance into a number of overlapping, fixed-length windows, with the window length equal to the exemplar size T. We then find a sparse representation for each window individually. Finally, depending on the noise robustness approach, the results for overlapping windows can be recombined and averaged.

Consider a noisy speech utterance $\mathbf{Y}_{\text{utt}}$ represented as a magnitude spectrogram of size $B \times T_{\text{utt}}$. We slide a window, a matrix of size $B \times T$, through $\mathbf{Y}_{\text{utt}}$, using a window shift of Δ frames. Thus, we obtain a sequence of windowed segments $\mathbf{Y}_1, \ldots, \mathbf{Y}_W$, where W is the number of windows in the utterance. A graphical representation of this process can be found in Fig. 2.

The ratio of Δ and T determines the degree to which subsequent windows overlap. Larger step sizes Δ reduce computational effort, but might decrease representational accuracy. Throughout this paper, we keep the window shift constant at $\Delta = 1$ frame.

At each window position w, the segment is reshaped into an observation vector $\mathbf{y}_w$ similarly as was done for speech and noise exemplars in section II-A. The index of the window position w ranges from 1 to $W = T_{\text{utt}} - T + 1$. The observation matrix $\boldsymbol{\Psi}$ of dimensions $E \times W$ has the observations vectors $\mathbf{y}_1, \ldots, \mathbf{y}_W$ as its columns.

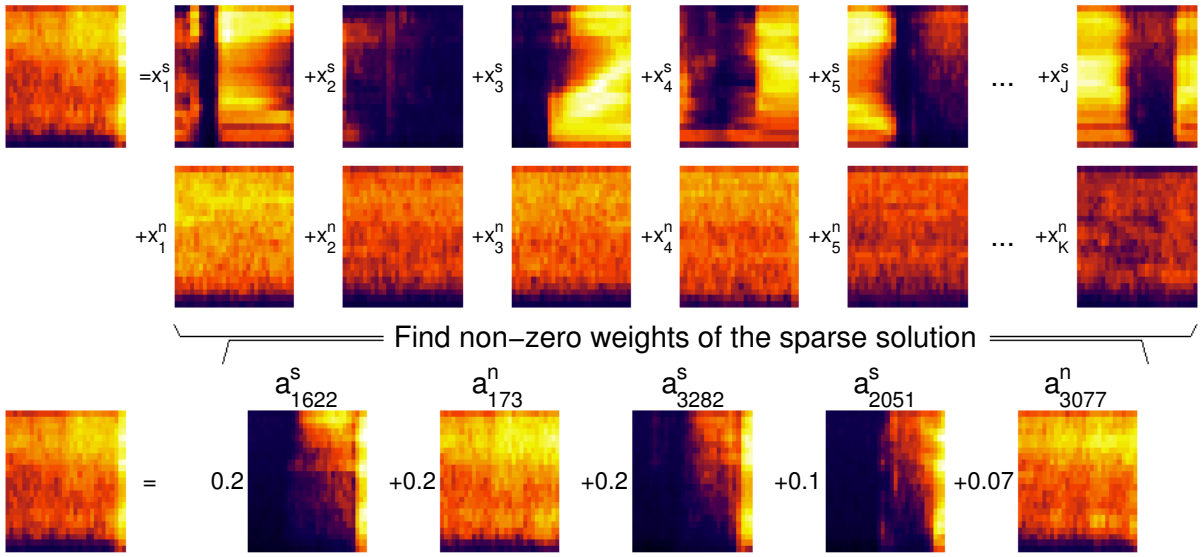

Fig. 1. An example of finding a sparse representation of a noisy speech signal as a linear combination of noise and speech exemplars. The speech and noise were mixed at SNR = 0 dB. The vertical axis of each spectrogram represents the frequency index, the horizontal axis time. The magnitude spectrograms are displayed using a logarithmic colour scale, with higher energies corresponding to brighter colours. The first row depicts the dictionary $\mathbf{A}^{\mathrm{s}}$ containing speech exemplars, while the second row corresponds to the dictionary $\mathbf{A}^{\mathrm{n}}$ containing noise exemplars. The third row shows the five largest non-zero activations of the sparse linear combination.

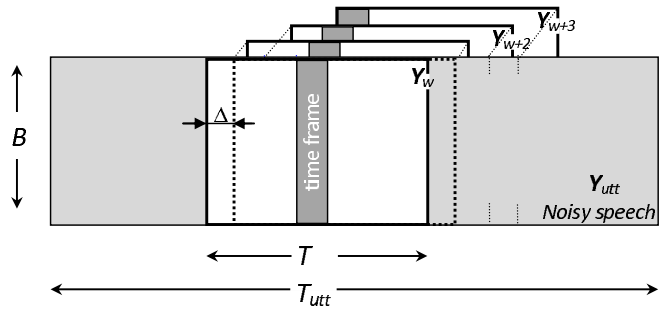

Fig. 2. Schematic diagram of time-continuous processing using overlapping windows.

Using the notation introduced above, we write the equivalent of (2) for the utterance $\mathbf{Y}_{\mathrm{utt}}$ with overlapping windows compactly as:

$$\mathbf{\Psi} \approx \mathbf{AX} \quad \text{s.t.} \quad \mathbf{X} \geq 0 \tag{9}$$

with the columns of matrix $\mathbf{X} = [\mathbf{x}_1 \;\; \mathbf{x}_2 \ldots \mathbf{x}_W]$ consisting of the sparse representations of each window. The activation matrix $\mathbf{X}$ of dimensions $L \times W$ now describes the activations of the exemplars for the entire utterance.

III. Sparse classification

In this section we describe the hybrid exemplar-based/HMM method called sparse classification. In our hybrid system we keep the topology of the HMM system unchanged, thus describing the speech structure in the conventional way in terms of a sequence of states. Rather than estimating the likelihoods of the states by means of GMMs, the calculation of likelihoods is based on the activations of exemplars described in II-A.

Sparse classification, first introduced in [20] for the classification of isolated digits, was extended to enable the recognition of connected digits without added noise [18]. In [19] it was shown that the method can be extended to noise robust connected digit recognition, while the method was further refined in [27].

A. Calculating speech state likelihoods

Assuming a state-level labelling of each frame in the speech data used to construct exemplars is available, we can label each frame $t = 1, \ldots, T$ in each speech exemplar $\mathbf{a}^{\mathrm{s}}_j$ with a state label $q_{j,t} \in [1, Q]$, where Q is the total number of states.

Using the frame-by-frame state labelling of the exemplars, we encode the labelling of each exemplar $\mathbf{a}^{\mathrm{s}}_j$ with label matrix $\mathcal{L}_j$. $\mathcal{L}_j$ is a sparse, binary matrix of dimensions $Q \times T$, the entries having values $[\mathcal{L}_j]_{q,t} = \delta(q, q_{j,t})$, where δ is the Kronecker delta function. The label matrix stores the temporal information of the states within an exemplar. Fig. 3 illustrates two examples of exemplars and their corresponding state label matrices.

Denoting the speech exemplar weights calculated for window w by $x^s_{w,j}$, $j = 1, \ldots, J$, we calculate state likelihood matrix $\mathbf{L}_w$ in window w as the weighted sum of exemplar label matrices as

$$\mathbf{L}_w = \sum_{j=1}^{J} \mathcal{L}_j x^s_{w,j} \tag{10}$$

The columns of $\mathbf{L}_w$ are denoted with vectors $\mathbf{l}_{w,t}$, $t = 1, \ldots, T$. State likelihood estimates of frames from overlapping windows are combined by summing the likelihoods of the frames of all the windows in which they occur, taking into account the exact temporal positions of the frames.

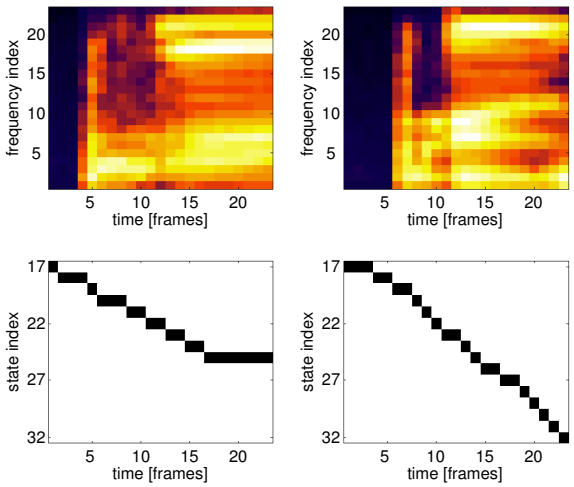

Fig. 3. Two exemplars of length $T = 25$ representing two different realisations of the beginning of the digit "two". The horizontal axes indicate the time in frames. The top panels illustrate the magnitude spectra, with a bright colour indicating a higher value. The lower panels illustrate the state label matrices explained in section III-A. The state indices $[17, 32]$ are the 16 states underlying the digit "two". The duration of the digit on the left exceeds 250 ms, and therefore the first 25 frames only cover part of the 16 states. The digit on the right takes less than 250 ms and thus covers all states.

The combined state likelihood vector $\mathbf{l}_\tau^{\text{utt}}$ for each frame $\tau = 1, \ldots, T_{\text{utt}}$ is given as

$$\mathbf{l}_\tau^{\text{utt}} = \sum_{t=\max(1,\tau-T_{\text{utt}}+T)}^{\min(T,\tau)} \mathbf{l}_{\tau-t+1,t} \tag{11}$$

After obtaining the state likelihoods for the entire utterance, we use the Viterbi algorithm to find the state sequence that maximises total likelihood.

B. Silence likelihoods

In the sparse classification approach, the likelihoods of silence states cannot be reliably estimated from noisy utterances using the method described above. As silence is absence of speech energy, a sparse representation of magnitude spectrograms models silence with all exemplar weights close or equal to zero. Since the state likelihoods are calculated by multiplication of the atom activations with the label matrix, silence is represented by all state likelihoods having a low value. After frame-wise normalisation of the likelihoods, the resulting likelihoods give rise to numerous insertion errors in areas where there was silence, for example at the begin and end of each utterance.

In the approach used in this work, we modify the speech and silence likelihoods produced by SC to circumvent this problem. In a nutshell, we first measure the activity of speech from the speech and noise exemplar activations. This measure is used to change the balance between the existing speech and silence state likelihoods, effectively boosting the silence likelihoods when there is no speech activity. The complete procedure, based on preliminary work described in [27], is given in Appendix B.

IV. Sparse representations for feature enhancement

As an alternative way to do noise robust ASR with the proposed exemplar-based framework, we use the sparse representations of speech and noise to estimate clean speech spectrograms, i.e., do feature enhancement. This is similar to source separation methods based on NMF [11]–[15], but the model is extended to deal with overlapping windows that span multiple frames.

Because of the use of a sliding window approach, the model (1) is obtained separately for each window $w = 1, \ldots, W$, each of which consists of T frames. By denoting the spectrum vector of the t-th frame of speech exemplar j by $\mathbf{a}_{j,t}^{\text{s}}$, the clean speech estimate $\tilde{\mathbf{s}}$ for the t-th frame of window w can be written as

$$\tilde{\mathbf{s}}_{w,t} = \sum_{j=1}^{J} \mathbf{a}_{j,t}^{\text{s}} x_{j,w}^{s}. \tag{12}$$

Likewise, the noise estimate $\tilde{\mathbf{n}}$ is given by

$$\tilde{\mathbf{n}}_{w,t} = \sum_{k=1}^{K} \mathbf{a}_{k,t}^{\text{n}} x_{k,w}^{\text{n}}. \tag{13}$$

For each frame $\tau = 1, \ldots, T_{\text{utt}}$ of the utterance, the models pertaining to overlapping windows are summed to obtain the speech and noise models. Normalisation by the number of overlapping windows is omitted, since it is cancelled by the later processing stages. When the window position within the utterance and the frame position within a window are taken into account, summation in frame τ results in the speech model

$$\hat{\mathbf{s}}_\tau = \sum_{t=\max(1,\tau-T_{\text{utt}}+T)}^{\min(T,\tau)} \tilde{\mathbf{s}}_{\tau-t+1,\tau} \tag{14}$$

Similarly, we add the noise spectra in overlapping windows to get

$$\hat{\mathbf{n}}_\tau = \sum_{t=\max(1,\tau-T_{\text{utt}}+T)}^{\min(T,\tau)} \tilde{\mathbf{n}}_{\tau-t+1,\tau}. \tag{15}$$

The resulting frame-wise estimates are grouped into speech and noise spectrogram utterance matrices:

$$\hat{\mathbf{S}}_{\text{utt}} = \left[\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_{T_{\text{utt}}}\right] \tag{16}$$

$$\hat{\mathbf{N}}_{\text{utt}} = \left[\hat{\mathbf{n}}_1, \ldots, \hat{\mathbf{n}}_{T_{\text{utt}}}\right] \tag{17}$$

The reconstructed speech spectra could be used directly as an estimate of clean speech features, as was done in NMF-based source separation with dictionary elements which span only a single frame [13]. In our exemplar-based framework, we obtain better results (in terms of recognition accuracy at high SNRs) by using a time-varying filter

$$\mathbf{h}_t = \hat{\mathbf{s}}_t ./ (\hat{\mathbf{s}}_t + \hat{\mathbf{n}}_t) \tag{18}$$

and calculating the enhanced features in each frame as $\mathbf{h}_t .* \mathbf{y}_t$. Unlike the reconstruction in (14), filtering the noisy spectra

takes the residual into account, that is, the noisy speech energy not modelled by the linear combination of speech and noise exemplars. The filtering approach is commonly used in source separation systems based on a linear model, for example in [14], [28], and [15]. This feature enhancement approach is also analogous to Wiener filtering in the frequency domain [29].

V. Sparse representations for missing data techniques

The missing data technique (MDT) approach to robust speech recognition [4], [5] is known for its high accuracy at high SNRs and its ability for dealing with non-stationary noise types. MDT is built on the assumption that one can estimate—prior to decoding—which spectro-temporal elements in the spectrogram are reliable (i.e., dominated by speech) and which are unreliable (i.e., dominated by background noise). The clean speech information in the unreliable features is considered *missing* and speech recognition must be done with partially observed data.

To do this, we employ the so-called *imputation* approach [30] which handles the missing features by replacing them with Gaussian-dependent clean speech estimates during decoding [31]. The difference between the feature enhancement approach described in section IV and imputation of the features, is that the latter is potentially more powerful because the Gaussian-dependent imputation approach can use information on the hypothesised state and digit identities.

The reliability estimates of noisy speech features are referred to as a *missing data mask*. We use the exemplar-based sparse representation framework to obtain a missing data mask $\mathbf{M}_{\text{utt}}$:

$$M_{\text{utt}}(b,\tau) = \begin{cases} 1 \stackrel{def}{=} \text{reliable} & \text{if } \frac{\hat{S}_{\text{utt}}(b,\tau)}{\hat{N}_{\text{utt}}(b,\tau)} > \theta \\ 0 \stackrel{def}{=} \text{unreliable} & \text{otherwise} \end{cases} \tag{19}$$

with the spectro-temporal magnitude speech and noise estimates $\hat{S}_{\text{utt}}$ and $\hat{N}_{\text{utt}}$ from (16) and (17), respectively. The constant θ is an empirically determined SNR threshold.

VI. Baseline recognisers

In this work, we compare the results obtained with the exemplar-based framework with two noise robust recognisers. The first is a multi-condition trained recogniser: a standard GMM-based recogniser trained on a mixture of clean and noisy speech of various noise types and SNRs. As an additional noise robustness measure mean and variance normalisation was used. On AURORA-2, multi-condition recognisers employing mean and variance normalisation are known to achieve state-of-the-art performance, often outperforming much more sophisticated noise robustness techniques [32].

As a second baseline, we use a MDT-based recogniser employing the so-called harmonicity missing data mask [25]. In the harmonicity mask, the noisy speech signal is first decomposed in a harmonic and a residual part using a least squares fitting method. The harmonic energy is then used as an estimator of the clean speech energy and the residual as an estimator for the noise energy, for use in (19).

VII. Experiments

In order to investigate the effectiveness of the exemplar-based framework, we compared the recognition accuracies of the various approaches described above using material from the AURORA-2 database. For each of the five methods (three exemplar-based methods, a baseline MDT-based recogniser and a multi-condition trained recogniser), we investigated word recognition accuracy as a function of SNR. For the exemplar-based approaches, we also investigated word recognition accuracy as a function of exemplar size.

A. Experimental setup

1) Recognition task: For our recognition experiments we used material from test set 'A' and 'B' of the AURORA-2 corpus [33]. The material we selected from test set A comprises 1 clean and 12 noisy subsets, with the noisy subsets containing four noise types (subway, car, babble, exhibition hall) at three SNR values, $15, 5$ and -5 dB. From test set B, which contains four different noise types (restaurant, street, airport, train station), we selected the same SNR subsets. Each subset contains 1001 utterances, with each utterance containing one to seven digits '0-9' or 'oh'. We evaluated word recognition accuracy by averaging the results over the four noise types.

The training material of AURORA-2 consists of a clean and a multi-condition training set, each containing 8440 utterances. The multi-condition training set was constructed by mixing the clean utterances with noise at various SNRs: $= \inf, 20, 15, 10, 5$ dB. The noises that were used originate from the same noise samples used to create test set A.

2) Finding sparse representations: Acoustic feature vectors used in the exemplar-based framework consisted of Mel frequency magnitude spectra: $B = 23$ frequency bands with centre frequencies starting at 100 Hz, using a Hamming window with a frame length of 25 ms and a frame shift of 10 ms.

The exemplar-based framework was implemented in MATLAB. The update rule (8) was run for 200 iterations which was enough to obtain solutions that had sufficiently converged. As in [19], the sparsity parameter was set to $\lambda = 0.65$ for speech exemplars and to $\lambda = 0$ for noise exemplars.

3) Dictionary creation: The speech and noise dictionaries were created in a two-step procedure which is repeated for each exemplar size $T \in \{5, 10, 20, 30\}$ frames. First, from each noisy utterance in the multi-condition training set two segments were selected of length T by choosing a random offset. The segments were allowed to overlap and no effort was made to exclude silence frames from the exemplars. For these segments, rather than using the noisy speech directly, the underlying clean speech and noise originally used for creating the noisy speech were extracted from their respective spectrograms and added to the speech and noise dictionaries. This resulted in initial speech and noise dictionaries consisting of 16880 exemplars.

In the second step, we created the speech dictionary by randomly selecting 4000 exemplars from the set of 16880. Experiments (not shown) revealed that for this dictionary size,

the choice of a random subset from the initial dictionary did not influence recognition significantly. For the initial noise dictionary, we first removed exemplars pertaining to silence (corresponding to clean speech utterances in the multi-condition train set) and then selected 4000 noise exemplars from the remaining 13504 exemplars. Dictionary creation took place only once, and dictionaries are kept fixed throughout all experiments.

4) Speech recognisers: Two speech recognisers were used. The first, used only for the multi-condition trained baseline recogniser, is the HTK-based recogniser described in [33], for which the configuration scripts are included in the AURORA-2 distribution. The only modifications that were made were the use of the zeroth cepstral coefficient in place of the log-energy and the use of per-utterance mean and variance normalisation of the cepstral features.

All other experiments make use of the second recogniser, a MATLAB implementation of the HMM-based missing data recogniser described in [31]. The acoustic models, trained on the clean speech in the training set, consist of 11 whole-word models with 16 states, as well as an additional 3-state silence word, resulting in a $Q = 179$ dimensional state-space. Every state was modelled by a mixture of 16 Gaussians with diagonal covariances.

The recogniser performs per-Gaussian-conditioned imputation during recognition, guided by a missing data mask. As input the decoder requires log-compressed Mel-band magnitude spectra and their first and second time derivatives. During decoding, the spectral features are converted to PROSPECT features, an alternative to cepstral features [31] which is computationally more efficient for missing data imputation. Missing data masks for the first and second time derivatives of the (static) speech features were calculated by taking the first and second time derivatives of the missing data mask pertaining to static features (cf. [34]). The way the recogniser was used differs for the Missing Data Technique (MDT), Feature enhancement (FE) and sparse classification (SC) experiments:

MDT The noisy magnitude features (described in Section VII-A2) were log-compressed, after which their first and second time derivatives were calculated. The missing data mask used for the missing data baseline is the harmonicity mask described in [25] with $10\log_{10}(\theta) = -9$ dB. For the missing data mask provided by our exemplar-based framework, we determined θ for each exemplar size separately by maximising recognition accuracy on the multi-condition training set. Based on these experiments, $10\log_{10}(\theta)$ was set to $\{-2, -1, 0, 0\}$ dB for exemplar sizes $T \in \{5, 10, 20, 30\}$, respectively.

FE The estimated clean speech magnitude features were log-compressed, after which their first and second time derivatives were calculated. As a missing data mask a mask was used that labels all features reliable, thus ensuring no imputation is done. The original clean speech models were updated by single-pass retraining the acoustic models on the enhanced spectra of the clean speech training data. The single-pass retraining consisted of re-estimating the Gaussian means and covariances on the processed (enhanced) speech using the original Gaussian-mixture weights and the canonical transcription.

SC The forced alignment of the clean speech training set with the canonical transcription, used for labelling the speech dictionary, was done using log-compressed magnitude features and their first and second time derivatives in combination with a missing data mask that labels all features reliable. During recognition, only the back-end of the recogniser was used in order to do Viterbi decoding.

B. Results

The speech recognition results from our experiments on AURORA-2 are displayed in Fig. 4. In each panel we display the results of the following methods:

- The multi-condition trained recogniser (M), described in [33], further augmented with mean and variance normalisation.
- The baseline MDT recogniser (I), described in [31].
- The missing data mask estimation (SMDT) approach with the mask derived from exemplar-based estimates of speech and noise, described in section V.
- The exemplar-based feature enhancement (FE) approach described in section IV.
- The exemplar-based sparse classification (SC) approach described in section III.

For clean speech, the top row in Fig. 4, we can observe that the SMDT and FE methods achieve similar recognition accuracies as the MDT-baseline recogniser at 99.3%. Moreover, there is no significant difference between the use of different exemplar sizes for the SMDT and FE exemplar-based methods, with the possible exception of $T = 5$. The multi-condition recogniser has an accuracy of 98.4% and the SC method achieves at most 96.6% accuracy at $T = 10$. For other exemplar sizes SC achieves lower accuracies.

In the left panel of the second row of Fig. 4, corresponding to test set A, SNR = 15 dB, we can observe that the multi-condition recogniser at 97.8% and the SMDT method at $T = 10$ with 97.4% now outperform the other methods by a small but significant margin. For all exemplar-based methods, we can observe clear differences in accuracy between the exemplar sizes, with the best performance being obtained with $T = 10$. In the right panel of that row, corresponding to test set B, we can observe a different result. While the multi-condition recogniser still performs best with 97.8% , SC now performs second best with 93.7% at $T = 20$, followed by FE with 91.3% accuracy at T=5.

At SNR = 5 dB, displayed in the third row, most exemplar based methods now perform better than the MDT-baseline. The SC method performs second best after the multi-condition recogniser, with 88.7% accuracy at $T = 20$ for test set A.

In the bottom row of Fig. 4, we can observe the results for SNR = -5dB. On test set A the SC method performs much better than all other methods, reaching 57.1% accuracy at $T = 30$. On test set B, the multi-condition recogniser performs best with 40.6% accuracy followed by SC at 37.0%.

All exemplar-based methods perform better than the MDT-baseline, although only by a small margin when using the SMDT method.

VIII. Discussion

A. Clean speech

In the results of our experiments on clean speech, we observed that SMDT and FE methods delivered the same high recognition accuracies as the MDT-baseline recogniser. The high accuracies of the MDT-based systems are due to the estimated masks correctly identifying all features as reliable. Consequently, no features are imputed and the recogniser (which is trained on clean speech) obtains the same results as would have been obtained without missing data imputation.

In the feature enhancement (FE) method, the reconstructions of speech and noise results in a *filter* (cf. (18)) that leaves the original clean speech features mostly unchanged. While small differences in the resulting features could result in an accuracy loss, this is compensated by the retraining of the acoustic models (cf. section VII-A4).

The multi-condition trained recogniser achieves lower accuracies on clean speech because its speech model is optimised for noisy speech rather than clean speech. Sparse classification (SC) also achieves lower accuracies, but for a different reason. SC provides state likelihoods and thus is dependent on a sparse representation of speech that uses exemplars with the 'correct' underlying state identities. There are three issues that might play in a role in the lower accuracies obtained with SC.

First, the exemplar dictionary is of limited size. Although exemplars from the dictionary can be linearly combined to describe the variation in the observations, the dictionary probably does not cover the entire acoustic space spanned by the complete training database. Increasing the size of the extracted dictionary can improve performance [18]. For maximising the performance, a more principled, possibly learning-based, method is probably needed to find a combination of exemplars which cover the entire training database. This topic is addressed in more detail in section IX-C.

The second issue could be the use of magnitude spectrum features. We tackled the problem of the large dynamic range of magnitude spectrum features by normalising the dynamic range of the features (cf. section II-B). In addition, we used the KL divergence rather than the Euclidean distance to give a more balanced weight to large and small magnitude values. However, it might be that even after these operations the distance measure neglects information in low-energy observations.

The final issue that might play a role is the presence of the noise dictionary when recognising clean speech. In clean speech, noise dictionary exemplars still get occasionally activated which can result in a less descriptive combination of speech exemplars. A principled approach to deal with this issue, as well as with the issue of silence balancing, would be to use machine learning techniques to learn the mapping from exemplar activations to likelihoods.

B. Noisy speech

In the presence of the additive background noise, all the exemplar-based techniques generally perform better than the MDT-baseline system. At the same time, the multi-condition recogniser employing mean and variance normalisation performs often better still, even at lower SNRs.

Of the exemplar-based methods, SMDT does worst, especially at lower SNRs. On test set B the SMDT approach often does not perform better than the MDT-baseline method. While the exemplar-based framework obtains a fairly accurate reconstruction of the clean speech spectra, the estimates of the noise spectra often contain residual speech. SMDT suffers from the presence of residual speech in the reconstructed noise spectra, because the reliable/unreliable classification depends on thresholding the difference between speech and noise spectra. The hard threshold makes the classification sensitive to small estimation errors. In addition, the empirically tuned threshold turns out to be dependent on the SNR level.

The sparse classification (SC) method clearly performs better than the other exemplar-based methods at SNRs < 15 dB. Especially at SNR -5 dB SC achieves much higher accuracies, up to 57.1% with $T = 30$ for test set A. It is clear that SC does so well on noisy speech because the underlying states (and therefore digit identities) are captured by the exemplar activations themselves. The fact that the enhancement methods and the missing data mask approach achieve lower accuracies, suggests that in noisy environments it is harder to estimate noise-free spectrograms than to directly estimate the underlying state or digit identity.

C. Performance differences between test sets A and B

The results reveal a distinct difference in performance between test set A and B: The performance of all exemplar-based methods is worse on test set B. The difference is due to the fact that the noise dictionary contains the same noise types as those found in test set A. Especially for the SMDT method, the performance of which critically depends on the accuracy of the noise estimate, this has a detrimental effect. Surprisingly, the multi-condition trained recogniser, which is also trained on noises encountered in test set A, seems to be the most noise-robust method on test set B. An explanation for this result is found in the fact that the noises in test set B, although originating from different noise sources, have a similar average spectral content as those found in test set A [33]. The noise exemplars employed in SC, on the other hand, also model the time structure of the noise and thus do not benefit from the similarity in average spectral content.

In order to study the influence of the noise match/mismatch in more detail, we performed an additional SC experiment on the full test set of AURORA-2 in which the real noise dictionary was replaced by a collection of completely artificial noise exemplars. These noise exemplars, first introduced in [35], consist of constant noise activity within a single frequency band for the duration of the exemplar. In other words, they are $B \times T$ dimensional all-zero matrices, with only one row (frequency band) having a non-zero, constant value. Thus, the total number of noise exemplars is only $B = 23$. In

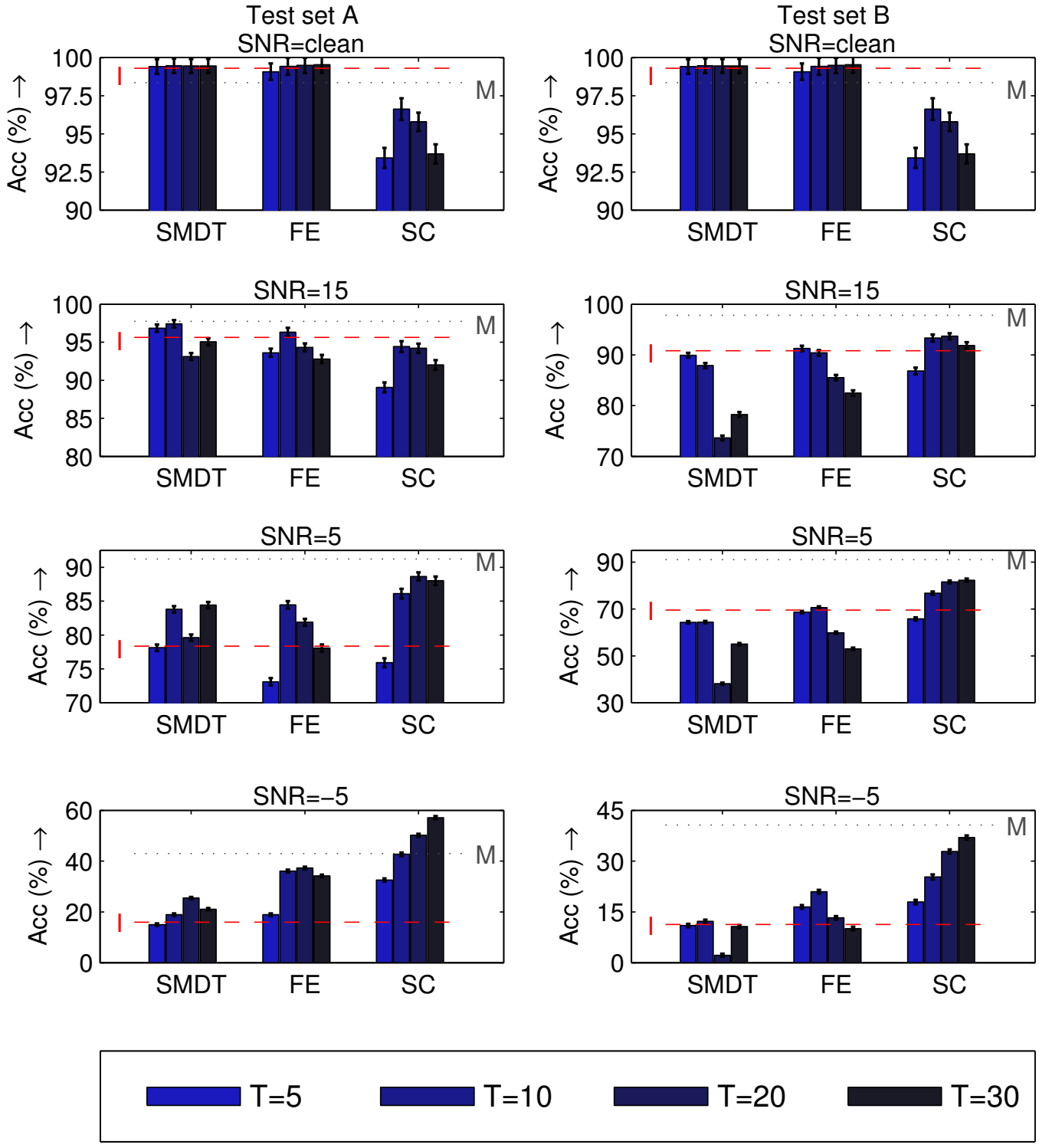

Fig. 4. Recognition results per test set and SNR. Recognition accuracy is displayed on the vertical axes. Vertical bars around the maxima indicate 95% confidence intervals. In each figure we display the results of missing data mask estimation (SMDT), feature enhancement (FE) and sparse classification (SC). For each exemplar-based method, we display four bar graphs, corresponding to the four exemplar sizes $T \in \{5, 10, 20, 30\}$. The baseline recogniser employing missing data imputation (I) and the multi-condition trained baseline recogniser (M) are indicated by dashed and dotted lines, respectively. The left panel corresponds to recognition on test set A and the right panel corresponds to recognition on test set B. The top row corresponds with clean speech, the second with noisy speech artificially corrupted with noise at 15 dB. The third and final row correspond to 5 and -5 dB. SNR, respectively. Note that the range of the vertical axis can differ between test sets A and B.

the experiment, all parameter settings and normalisation steps were kept the same as when using a real noise dictionary (cf. Sections II-A and VII-A).

In Fig. 5 the results of SC ($T = 30$) with both real and artificial exemplars are shown, as well as the performance of the MDT baseline and multi-condition baseline recognisers. The results show a decrease in the accuracy on test set A, but an *increase* in accuracy on test set B. The decrease in accuracy on test set A might be an indication that the time structure of the noise types in test set A is more difficult to model with the artificial exemplars. The fact that the SC performance on set B increased, is another indication that in some sense the real noise exemplars model are overfitting the noise, resulting in a lack of generalisation.

Still, the results of this small experiment also show that while SC benefited from the match between noise dictionary and noises encountered in the noisy speech of test set A, the method can also provide noise robustness with —by definition mismatched— artificial noise exemplars.

D. Effectiveness of Sparse Classification

From the discussions above it is clear that in its current form, SC does not yet reach high enough accuracies to be a replacement for GMM-based calculation of likelihoods, such as those that can be obtained with a multi-condition trained recogniser. At the same time, the results at low SNRs show the SC approach has potential, especially when there is some knowledge of the corrupting noises that can be present. The fact that at lower SNRs, it is harder to estimate noise-free spectrograms than to directly estimate the underlying state or digit identity, makes a strong case for using the SC-based likelihoods scores to improve noise robustness.

The SC framework presented in this work is, to the best of our knowledge, the first instance of using exemplar-based techniques for noise robust ASR. As such, it may serve as a starting point for exemplar-based ASR research. Being an exemplar-based technique, SC is quite different, and potentially more flexible, than model-based noise robustness methods.

In popular model-based compensation techniques such as parallel model combination (PMC) and Vector Taylor Series (VTS) approaches [3], [36], [37], the acoustic model is updated in each frame to account for the noise estimated to be present in that frame. In the end, however, the acoustic model still describes noisy speech, which can lead in a lack of discriminative power at lower SNRs. In the SC approach the noise and speech are separately modelled. The fact that the SC algorithm has the freedom of choosing any noise exemplar at any point in the utterance, without relying on any initial noise estimate, can make it more flexible. After all, any additional information on the noises that might be encountered in the utterance can simply be added to the noise dictionary on the fly.

Also, it is important to notice that the unfavourable comparison to the multi-condition baseline recogniser may very well be due to its lower performance on clean speech: It is likely that these lower accuracies propagated into the lower SNRs. Thus, improving the results at high SNRs may make SC a more viable alternative. In Section IX we discuss various options for improving the results. Preliminary research on one of the options discussed there, combination of SC and GMM-based likelihoods, has already revealed that 98.8% accuracy on clean speech is easily obtainable [38].

E. Influence of using time-context

One characteristic of the exemplar-based representations presented in this work, is that they allow straightforward modelling of multiple frames of time-context. From the results, we can generally observe that longer exemplar sizes ($T = 20, 30$) improve performance at lower SNRs, while at higher SNRs shorter exemplar sizes ($T = 5, 10$) work better. The SMDT method is an exception, however, when considering its performance on test set B. The reason for this is probably that its performance, especially with larger exemplar sizes, is critically dependent on the exemplar size-dependent threshold value. These thresholds, tuned on the multi-condition set which contains the same noise types used in test set A, do not seem to generalise well, which causes shorter exemplar sizes to do better.

The reason for the better performance of longer exemplars at low SNR is that including more time context prevents confusion with noise exemplars, by imposing more constraints on the search for a sparse linear combination of exemplars. A similar result was found in [39], in which log-spectral exemplar-based representation were used for missing data imputation.

At the same time, using larger exemplars at higher SNRs decreases the accuracy because it becomes more difficult to accurately describe clean speech as a linear combination of such large exemplars. In [18] it was shown for clean speech, that for larger exemplar sizes a larger dictionary is needed to reach the same accuracies.

Another downside of using more time-context may be the more accurate modelling of noise. As pointed out in Section VIII-C, modelling the time-context of noise may reduce performance when there is a mismatch between the noise types in the dictionary and those encountered in the noisy speech. While arguable not a problem for source separation scenarios in which the characteristics of both sources are known, this is detrimental for noise robust ASR in which the corrupting noise type is often difficult to predict in advance.

We conclude that the use of exemplars spanning multiple frames of time-context is beneficial if the underlying sources are known. Based on these findings, it might be beneficial to combine multiple exemplar sizes into a single system: using smaller exemplar sizes to improve generalisation while longer exemplars more accurately model the time structure of known sources.

F. Model complexity and computational effort

Because the exemplar-based methods use a dictionary that contains a large number of real speech spectrograms, the number of parameters needed to model the speech is larger than for the corresponding GMM model. In the GMM model, each GMM state is represented as a mixture of B-dimensional

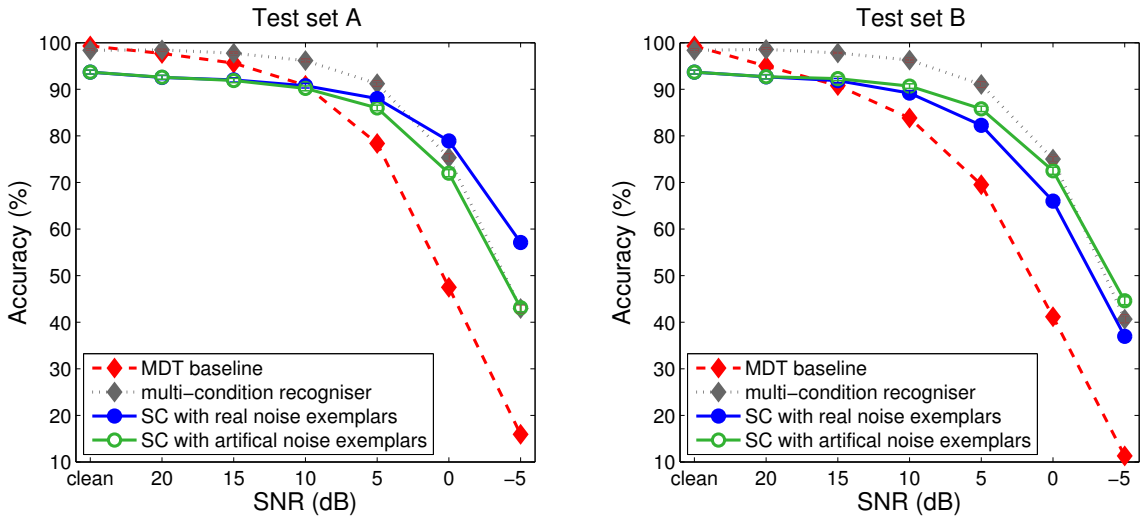

Fig. 5. Recognition results per test set as a function of SNR. Recognition accuracy is displayed on the vertical axes. Vertical bars around the maxima indicate confidence intervals at a 95% confidence. In each figure we display the results of sparse classification using real noise exemplars, sparse classification using artificial noise exemplars and the multi-condition recogniser. For the sparse classification methods, the exemplar size was $T = 30$.

mean and variance vectors for each Gaussian, resulting in $Q \times B \times G \times 2$ parameters, where G is the number of Gaussians per state. In the recognisers used in this work, which also model the first and second time derivatives of the static features, this amounts to a model size of approximately 3.95×10^5 parameters. The dictionary of the exemplar-based methods contains a total of $L \times B \times T$ entries. This means the clean speech is modelled by approximately 4.6×10^5 to 2.76×10^6 parameters, depending on the exemplar size.

To roughly characterise the computational effort needed, we did a test of the running time of the SC and MDT-baseline method on a single core of a 64-bit machine with a Core 2 Quad Q6600 2.4 GHz processor. We tested the utterance 'MIP_68385A', taken from test set A, subway noise type, SNR -5 dB, which has a length of 182 frames (1.82 seconds of speech). For the baseline method, the average running time to obtain noise robust likelihoods was 8.6 seconds, of which 4.9 seconds were spent on missing data mask estimation, and 3.5 seconds were spent on imputation and Gaussian evaluation. The running times for the SC method are given in Table I. The average time spent on Viterbi decoding was a negligible 0.08 seconds.

From Table I we can observe that, depending on the exemplar size T, the SC method is a factor 5 to 15 slower than the baseline MDT method. At the same time we observe that the running time is completely dominated by the time spent on minimising update rule (8). This algorithm, fortunately, lends itself well to parallelisation and speedups of a factor 30 and higher using modern graphics cards have already been reported for similar problems [40].

TABLE I
AVERAGE RUNNING TIMES FOR DIFFERENT EXEMPLAR SIZES. THE 'SOLVER' COLUMN DEPICTS THE TIME SPENT ON MINIMISING UPDATE RULE (8).

T	W	solver [s]	total [s]
5	178	42.4	42.5
10	173	60.3	60.4
20	163	100.1	100.4
30	153	129.2	129.5

IX. Future improvements and extensions

A. Finding a better linear combination of exemplars

The successful application of the sparse representation-based methods hinges on the success with which a correct linear combination of exemplars can be found. That success, in turn, is largely dependent on the constraints placed on the minimisation in (6). Aside from the constraints currently used in SC, e.g. non-negativity, sparsity, and the use of exemplars that span multiple frames, there are several options for additional constraints.

While the current framework models time-context by using exemplars that span multiple frames, we do not explicitly model the fact that in addition to the absolute energy levels, modulations or changes in the speech energy carry also important information. In HMM-based recognisers, this is typically modelled by using derivative features. Since derivative features are simply a linear combination of the static features, they could be combined into the current framework by splitting them into a negative and positive part and stacking them with the current features. Such an approach has already been used in NMF-based source separation [41].

Another approach to add additional constraints would be to take the identity of the exemplars into account by using 'group sparsity' [42]. At any window position, only a limited number of digits (usually only one) can be present. The linear combination of speech exemplars, while sparse, may still consist, however, of exemplars pertaining to many more different digit identities. Rather than enforcing the sparsity of the linear combination itself, we can enforce the linear combination to be sparse between groups (different realisations of the same digit). This could increase the accuracy of the resulting sparse representation and would also result in sparser

and thus better defined likelihoods.

As pointed out in Section VIII-A, another factor determining the success with which a sparse representation is found is the feature representation. In this work we employ magnitude domain features which allow for a simple formulation of additive noise and speech as a single sparse representation. As an alternative, it may be possible to allow speech exemplars to combine in the logarithmic domain, while at the same time allowing noise and speech to add in the magnitude domain. This will make cost function (6) more complex, however, and more research is needed to effectively minimise it.

B. Hybrid recognition using sparse classification

Aside from improving the underlying framework, a practical approach to improve the results of sparse classification, especially at high SNRs, would be to combine SC with conventional speech recognisers. Such a hybrid approach can take many forms. Using a 'tandem' approach [43], one could treat the likelihoods produced by SC as input features for a conventional GMM based recogniser, in order to 'learn' the proper mapping between the internal state representation of the HMM-based recogniser and the exemplar activations underlying SC. Such an approach would have the advantage that we can employ any low-dimensional representation of the exemplar activations, such as phone-based labelling.

A more principled approach would be to combine the information provided by the SC framework with those provided by GMMs in a Dynamic Bayesian Network (DBN). In such a model, we can describe the joint probability of the GMM and SC likelihoods. Preliminary experiments on AURORA-2 with this approach, described in [38], showed that the DBN can have both the high performance at lower SNRs of SC while retaining the high performance at high SNRs as obtained by a GMM operating on cepstral features.

C. Large vocabulary speech recognition

The exemplar-based framework described in this work can be applied to large vocabulary speech without further adaptation, provided the speech can be represented as a sparse linear combination of exemplars. In [39] the sparsity of an exemplar-based representation on large vocabulary speech was investigated. It was shown that a dictionary consisting of randomly extracted exemplars can be used to sparsely represent large vocabulary spontaneous speech using no more than 30 exemplars at a time. Although in that work a logarithmic compression of the magnitude features was used (in combination with a Euclidean distance measure), the result gives confidence that at least in principle, exemplar-based sparse representations can be used for the representation of large vocabulary speech.

Still, a more principled approach toward the creation of an exemplar dictionary is probably required. An ideal exemplar-based speech dictionary should probably cover the full range of variation in speech phenomena. Random selection of exemplars ensures a good representation of the relative occurrence of speech phenomena, but in larger vocabulary tasks rare phones and pronunciations will easily be under-represented. Several alternatives are possible, such as ensuring that exemplars are selected from certain phonetic (or state) groups, clustering-based approaches or dictionary learning approaches [44].

Another issue is that in its current form, many exemplars are needed to provide shift invariance and cover variability in duration. An alternative for this is to modify the model to allow shift and duration invariance. Algorithms that can estimate the activations in such a model can be based on the convolutive NMF approaches described in [45] and [46].

X. Conclusions

We proposed the use of an exemplar-based framework in which noisy speech is modelled by a sparse linear combination of speech and noise exemplars. These exemplars consist of segments of speech or noise that span multiple time-frames, typically 50 to 300 ms. We proposed the use of the sparse classification method that uses such sparse representations to do hybrid exemplar-based/HMM decoding. The weights of the linear combination together with an HMM-state based labelling of the speech exemplars is used to feed noise robust HMM-state likelihoods to the Viterbi back-end of a conventional recogniser. Moreover, we described how the exemplar-based approach can be used as a source separation technique in order to do missing data mask estimation and feature enhancement.

We compared the sparse classification approach with the other exemplar-based approaches to noise robust recognition as well as a missing data based noise robust baseline recogniser and a multi-condition trained baseline recogniser. Results on the AURORA-2 database revealed that the sparse classification method outperformed the other exemplar-based methods at SNRs < 15 dB, achieving up to 57.1% accuracy at SNR= -5 dB. From this we concluded that at low SNRs, it is better to directly estimate the underlying state or digit identities from the sparse representation than to try to reconstruct a clean speech spectrogram.

When investigating the influence of using exemplars that include multiple frames of time-context, we found that in general, longer exemplars work better at lower SNRs. We concluded that the use of longer time-context is beneficial if the underlying sources are known, but that smaller exemplar sizes may be more effective for generalisation to unknown sources.

In comparison to the baseline recognisers, it was found that sparse classification only performed better at low SNRs. We discussed the various reasons for its lower performance on clean speech and outlined several promising ways for improving the performance of exemplar-based sparse representation methods in general, and sparse classification in particular. Future research is needed to establish the effectiveness of these strategies.

Appendix A
Derivation of the update rule

This appendix describes the derivation of update rule (8), used for minimisation of the cost function (6). First, we rewrite Eq. (6) as:

$$d(\mathbf{y}, \mathbf{A}\mathbf{x}) + d(\mathbf{0}, \mathrm{diag}(\boldsymbol{\lambda})\mathbf{x}) \tag{20}$$

where d is the KL divergence (7), $\mathbf{0}$ is an all-zero column vector of length E and $\mathrm{diag}(\boldsymbol{\lambda})$ is a diagonal matrix having the elements of $\boldsymbol{\lambda}$ in the diagonal. The sum (20) can be written as $d(\mathbf{z}, \mathbf{Z}\mathbf{x})$, with $\mathbf{z}^{\mathrm{T}} = \left[\mathbf{y}^{\mathrm{T}}\mathbf{0}^{\mathrm{T}}\right]$ and $\mathbf{Z}^{\mathrm{T}} = \left[\mathbf{A}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\lambda})\right]$.

For a function of this form an update rule was proposed in [9]:

$$\mathbf{x} \leftarrow \mathbf{x}. * (\mathbf{Z}^{\mathrm{T}}(\mathbf{z}./(\mathbf{Z}\mathbf{x})))./(\mathbf{Z}^{\mathrm{T}}\mathbf{1}). \tag{21}$$

By substituting $\mathbf{z}^{\mathrm{T}} = \left[\mathbf{y}^{\mathrm{T}}\mathbf{0}^{\mathrm{T}}\right]$ and $\mathbf{Z}^{\mathrm{T}} = \left[\mathbf{A}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\lambda})\right]$ in (21) we obtain update rule (8).

This update rule leads to non-increasing values of the cost function (20) as proven in [9]. While theoretically that does not ensure convergence to a stationary point [47], in our studies with real data we have observed that the algorithm converges sufficiently robustly.

APPENDIX B
SILENCE BALANCING

This appendix describes a method for modifying the original, potentially unreliable (cf. section III-B) ratio between speech and silence states. This is done through the use of a speech level estimate. This *speech activity* is calculated from the sum of speech exemplar activations per window:

$$r_w = \sum_{j=1}^{J} x^s_{w,j} \tag{22}$$

with r_w the speech activity in window w. A frame level estimate r_t is acquired by linear interpolation of r_w. r_t is then normalised to the $[0, 1]$ range over the complete utterance, with 0 denoting silence and 1 the maximum observed speech level.

In order to obtain a steeper division between speech and silence regions, we then calculate the *adjusted speech activity* $\hat{r}_t$ by applying a shifted and scaled logistic function:

$$\hat{r}_t = \frac{1}{1 + \exp(-\alpha r_t - \beta)} \tag{23}$$

with the threshold level and steepness controlled by the parameters α and β.

At each frame, speech state likelihoods are multiplied by a single scalar so that their sum equals $\hat{r}_t$. Similarly, silence states are scaled together so that their sum equals $1 - \hat{r}_t$. Consequently, the original ratio between speech and silence likelihoods is replaced by one defined by $\hat{r}$, and the likelihoods in each frame sum to unity.

In practice, it was found that the parameters α and β of the logistic function (23) are more stable when defined by a weight factor ϕ, which describes the overall influence of speech states, and a width factor χ, which represents the steepness:

$$\alpha = \frac{1}{\chi} \tag{24}$$

$$\beta = \log \frac{1 - e^{\phi\alpha}}{e^{\phi\alpha} - e^{\alpha}}. \tag{25}$$

TABLE II
SILENCE BALANCING PARAMETERS FOR DIFFERENT EXEMPLAR SIZES.

T	c_χ	c_0	c_ϕ
5	0.01	0.998	0.03
10	0.05	0.996	0.12
20	0.08	0.992	0.26
30	0.105	0.988	0.225

The parameters ϕ and χ were made exemplar size and SNR dependent. The SNR of each utterance is estimated as the ratio of the sum of speech and noise exemplar activations:

$$\mathrm{SNR} = \frac{\sum_{w=1}^{W}\sum_{j=1}^{J} x^s_{w,j}}{\sum_{w=1}^{W}\sum_{k=1}^{K} x^{\mathrm{n}}_{w,k}} \tag{26}$$

with the SNR estimate truncated to the $[0.04, 4]$ range. The parameters ϕ and χ are now calculated as:

$$\chi = c_\chi * \mathrm{SNR}^{-\frac{1}{2}}, \tag{27}$$

$$\phi = c_0 - c_\phi * \chi. \tag{28}$$

The constants c_χ, c_0, c_ϕ were optimised by maximizing the recognition accuracy on the multi-condition training set (cf. section VII-A1) for each exemplar size T separately using a grid search, and are given in Table II.

ACKNOWLEDGMENTS

The authors would like to thank Bert Cranen and Lou Boves for their help with the manuscript.

REFERENCES

[1] H. Bourlard, H. Hermansky, and N. Morgan, "Towards increasing speech recognition error rates," *Speech Communication*, vol. 18, pp. 205–231, 1996.

[2] P. Moreno, B. Raj, and R. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. International Conference on Audio, Speech and Signal Processing*, Atlanta, USA, 1996.

[3] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, 1996.

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

[5] M. Cooke, P. Green, and M. Crawford, "Handling missing data in speech recognition," in *Proc. International Conference on Speech and Language Processing*, 1994, pp. 1555–1558.

[6] N. S. Kim, D. K. Kim, and S. R. Kim, "Application of sequential estimation to time-varying environment compensation in speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 389 –395.

[7] B. J. Frey, L. Deng, A. Acero, and T. Kristjansson, "ALGONQUIN: Iterating laplace's method to remove multiple types of acoustic distortion for robust speech recognition," in *Proc. EUROSPEECH*, 2001, pp. 901–904.

[8] K. Yao and S. Nakamura, "Sequential noise compensation by sequential Monte Carlo method," in *Proc. Neural Information Processing Systems*, 2002, pp. 1205–1212.

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Information Processing Systems*, April 2001, pp. 556–562.

[10] E. J. Candés, , and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[11] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[12] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[13] 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.

[14] T. Virtanen and A. T. Cemgil, "Mixtures of gamma priors for non-negative matrix factorization based speech separation," in *Proc. ICA*, 2009.

[15] B. Raj, T. Virtanen, S. Chaudhure, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proc. International Conference on Speech and Language Processing*, 2010.

[16] 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, February 2009.

[17] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2010.

[18] J. F. Gemmeke, L. ten Bosch, L.Boves, and B. Cranen, "Using sparse representations for exemplar based continuous digit recognition," in *Proc. EUSIPCO*, Glasgow, Scotland, August 24–28 2009, pp. 1755–1759.

[19] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2010.

[20] J. F. Gemmeke and B. Cranen, "Noise robust digit recognition using sparse representations," in *Proc. of ISCA 2008 ITRW "Speech Analysis and Processing for knowledge discovery"*, 2008.

[21] G. S. V. S. Sivaram, S. K. Nemala, M. Elhilali, T. D. Tran, and H. Hermansky, "Sparse coding for speech recognition," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2010.

[22] J. F. Gemmeke, H. Van hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 272–287, 2010.

[23] M. D. Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template based continuous speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1377–1390, 2007.

[24] M. Van Segbroeck and H. Van hamme, "Applying non-negative matrix factorization on time-frequency reassignment spectra for missing data mask estimation," in *Proc. INTERSPEECH*, Brighton, UK, September 6–10 2009.

[25] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. International Conference on Audio, Speech and Signal Processing*, vol. 1, 2004, pp. 213–216.

[26] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2008.

[27] T. Virtanen, J. F. Gemmeke, and A. Hurmalainen, "State-based labelling for a sparse representation of speech and its application to robust speech recognition," in *Proc. Interspeech*, 2010, pp. 893–896.

[28] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Proc. Neural Information Processing Systems*, 2009.

[29] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, 2006.

[30] B. Raj, R. Singh, and R. Stern, "Inference of missing spectrographic features for robust automatic speech recognition," in *Proc. International Conference on Speech and Language Processing*, Sydney, Australia, November 30–December 4 1998, pp. 1491–1494.

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

[32] C.-P. Chen and J. A. Bilmes, "MVA processing of speech features," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 257 –270, jan. 2007.

[33] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA Tutorial and Research Workshop ASR2000*, Paris, France, September 18–20 2000, pp. 181–188.

[34] H. Van hamme, "Handling time-derivative features in a missing data framework for robust automatic speech recognition," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2006.

[35] J. F. Gemmeke and T. Virtanen, "Artificial and online acquired noise dictionaries for noise robust ASR," in *Proc. Interspeech*, 2010, pp. 2082–2085.

[36] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector taylor series for noise speech recognition," in *Proc. the International Conference on Spoken Language Processing*, Beijing, China, 2000.

[37] R. C. van Dalen, F. Flego, and M. J. F. Gales, "Transforming features to compensate speech recogniser models for noise," in *Proc. INTERSPEECH*, Brighton, UK, September 6–10 2009, pp. 2499–2502.

[38] Y. Sun, J. F. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves, "Using a DBN to integrate Sparse Classification and GMM-based ASR," in *Proc. Interspeech*, 2010, pp. 2098–2101.

[39] J. F. Gemmeke, B. Cranen, and U. Remes, "Sparse imputation for large vocabulary noise robust ASR," *Computer Speech & Language*, vol. In Press, Corrected Proof, 2010. [Online]. Available: DOI:10.1016/j.csl.2010.06.004

[40] P. Nagesh, R. Gowda, and B. Li, "Fast GPU implementation of large scale dictionary and sparse representation based vision problems," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2010, pp. 1570–1573.

[41] M. Van Segboeck and H. Van hamme, "Unsupervised learning of time-frequency patches as a noise-robust representation of speech," *Speech Communication*, vol. 51, no. 11, 2009.

[42] A. Majumdar and R. K. Ward, "Classification via group sparsity promoting regularization," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2009, pp. 861–864.

[43] S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky, "Feature extraction using non-linear transformation for robust speech recognition on the Aurora database," in *Proc. International Conference on Audio, Speech and Signal Processing*, 2000, pp. 1117–1120.

[44] 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.

[45] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

[46] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, 2007.

[47] C.-J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Transactions on Neural Networks*, 2007.