Coupled Dictionaries for Exemplar-based Speech Enhancement and Automatic Speech Recognition

Deepak Baby, *Student Member, IEEE,* Tuomas Virtanen, *Member, IEEE,* Jort F. Gemmeke, *Member, IEEE,* and Hugo Van hamme, *Senior Member, IEEE*

***Abstract*—Exemplar-based speech enhancement systems work by decomposing the noisy speech as a weighted sum of speech and noise exemplars stored in a dictionary, and use the resulting speech and noise estimates to obtain a time-varying filter in the full-resolution frequency domain to enhance the noisy speech. To obtain the decomposition, exemplars sampled in lower dimensional spaces are preferred over the full-resolution frequency domain for their reduced computational complexity and the ability to better generalize to unseen cases. But the resulting filter may be sub-optimal as the mapping of the obtained speech and noise estimates to the full-resolution frequency domain yields a low-rank approximation. This paper proposes an efficient way to directly compute the full-resolution frequency estimates of speech and noise using coupled dictionaries: an input dictionary containing atoms from the desired exemplar space to obtain the decomposition and a coupled output dictionary containing exemplars from the full-resolution frequency domain. We also introduce modulation spectrogram features for the exemplar-based tasks using this approach. The proposed system was evaluated for various choices of input exemplars and yielded improved speech enhancement performances on the AURORA-2 and AURORA-4 databases. We further show that the proposed approach also results in improved word error rates (WERs) for the speech recognition tasks using HMM-GMM and deep-neural network (DNN) based systems.**

***Index Terms*—Exemplar-based, noise robust automatic speech recognition, non-negative sparse coding, modulation envelope**

I. Introduction

SPEECH recordings taken from realistic environments typically contain degradations along with the required speech signal which reduce its intelligibility and also result in poor performance of speech related tasks like automatic speech recognition (ASR), automatic voice assistance, etc. Therefore, some speech enhancement mechanism is deployed as the first step in most of these applications to circumvent the degradations which are mainly introduced by the background noise and room reverberation.

In scenarios where a model for speech and noise is not known *a priori*, unsupervised techniques like spectral subtraction [1], Kalman filtering [2], using the periodic structure in speech [3], etc., have been successfully used for speech enhancement. But most of these approaches rely on stationarity assumptions on the noise, which are often invalid for realistic data. Alternatively, supervised techniques can yield improved performance using codebook based [4] or model based [5] approaches, since the models for speech and noise are known *a priori*.

In this work, we investigate speech enhancement on a single channel noisy recording in the presence of additive noise using non-negative matrix factorization (NMF) algorithms. Ever since its introduction [6], NMF has been successfully used for numerous source separation problems [7]–[9]. Given a dictionary containing atoms representing the sources, NMF-based algorithms decompose a noisy observation as a sparse non-negative linear combination of the atoms. In our framework, the atoms used are time-frequency representations of the training speech and noise data. The NMF-based decomposition thus yields estimates of speech and noise in the observation which can then be used to obtain a time-varying filter in the full-resolution frequency domain for speech enhancement.

One of the popular approaches in NMF-based algorithms is to use overcomplete dictionaries created using "exemplars" of speech and noise that are the directly sampled versions of the training speech and noise data itself [10]–[12]. Another approach is to train the dictionary atoms from the training samples using the NMF updates [13], where generalisable models for speech and noise are learned as undercomplete dictionaries [14,15]. A study presented in [16] compares these two approaches and showed that the NMF-learned dictionaries outperform the exemplar-based dictionaries for speech enhancement in reverberant environments. However, the comparisons are done only with undercomplete dictionaries. It is also observed that, given enough training data to create overcomplete dictionaries, using exemplars from the training data as such leads to better separation performance than the NMF-learned dictionaries [17,18]. In this work we use overcomplete dictionaries where exemplars are expected to work better and we refer this approach to as "exemplar-based approach".

The performance of an exemplar-based approach depends on two key factors: First, on how well the speech and noise can be differentiated in the chosen time-frequency representation or the "exemplar space". Popular choices of exemplar spaces include Mel-integrated magnitude spectra [10], DFT (refers to the magnitude of the short-time Fourier transform in this paper) [19] and Gabor filterbank coefficients [11]. Using DFT as the exemplar space has the advantage that the time-varying filter can be directly obtained in the full-resolution frequency (DFT) domain. However, such systems suffer from increased computational complexity, poor speech and noise separation especially in presence of babble noise [20] and inability to generalise well for unseen noise cases [21]. It is observed that using lower dimensional features like the Mel features can address most of these issues fairly well [21] and this introduces the second factor: how well we can map the resulting lower-dimensional estimates to the DFT space to obtain the time-varying filter? Most of the current approaches make use of a pseudo-inverse [12] to obtain the mapping which always yield a low-rank approximation of the estimates, resulting in a sub-optimal filter which cannot account for all the added noise content and results in poorer noise suppression.

In this work, we have three main goals. First, to effectively utilize the advantages of the low-dimensional features and to address the low-rank approximation, we propose to use coupled dictionaries, which has been used earlier to increase the spectro-temporal resolution [18,22], voice conversion [23] and dimensionality reduction for multi-label learning [24]. In this work, we make use of two (coupled) dictionaries: an input dictionary containing atoms sampled in the exemplar space where the NMF-based decomposition is to be done and a coupled output dictionary containing the corresponding DFT

D. Baby, J. F. Gemmeke and H. Van hamme are with the Speech Processing Research Group, Electrical Engineering Department (ESAT), KU Leuven, 3000 Leuven, Belgium (e-mail: Deepak.Baby@esat.kuleuven.be; jgemmeke@amadana.nl; Hugo.Vanhamme@esat.kuleuven.be).

T. Virtanen is with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland (email: Tuomas.Virtanen@tut.fi).

This work has been funded with support from the European Commission under Contract FP7-PEOPLE-2011-290000 (INSPIRE) and IWT-SBO Project 100049 (ALADIN).

exemplars to directly reconstruct the estimates in the DFT domain. This approach thus can obtain a better decomposition at a reduced computational complexity, and make use of the resulting weights or *activations* of the input dictionary atoms to directly reconstruct the DFT estimates using the coupled output dictionary, which will be explained in Section II.

Second, we introduce using modulation spectrogram (MS) features [25] for exemplar-based speech enhancement. The MS representation for speech was introduced as part of a computational model for human hearing and a better separation between speech and noise can be expected in the MS domain considering the fact that speech and noise often have different modulation frequency contents. However, obtaining the MS representation involves non-linear operations which makes it hard to invert to the frequency domain where the mixture signal is processed. In this work, we investigate the use of coupled dictionaries to reconstruct the underlying DFT features following the decomposition in the MS domain for exemplar-based speech enhancement and ASR tasks.

Finally, we investigate the performance of various state-of-the-art automatic speech recognition (ASR) tasks on these enhanced speech data. ASR evaluation serves two purposes in this work. First, the recognition performance acts as an additional evaluation measure to assess the utility of the enhanced speech data on small and large vocabulary speech recognition. Second, we investigate how much the HMM-GMM based and deep-neural network (DNN) based state-of-the-art ASR systems can benefit from making use of the enhanced data.

The rest of the paper is organized as follows: Section II details the proposed exemplar-based speech enhancement technique using coupled dictionaries. The various choices of input exemplars investigated in this work are described in Section III. The evaluation setup is explained in Section IV followed by some results and observations made on the experiments done on the AURORA-2 database in Section V. Section VI details the results obtained for speech enhancement and ASR evaluations on the AURORA-4 database. Section VII concludes the paper along with some directions for future work.

II. Speech Enhancement using Coupled Dictionaries

A. Compositional model for noisy speech using NMF

Exemplar-based separation of speech and noise in a noisy recording makes use of a speech dictionary $\mathbf{A_s}$ containing J exemplars sampled from segments of clean speech and a noise dictionary $\mathbf{A_n}$ containing K exemplars sampled from segments of noise that corrupt speech. Exemplars are spectro-temporal representations of the training data, with the spectral axis referred to as frequency bins or coefficients and temporal axis as frames. The principle behind the approach is that the noisy speech, being an addition of speech and noise, can be approximated as a weighted sum of atoms in the speech and noise dictionaries. The exemplars may span multiple, say T, frames (which are reshaped to a vector) to capture the temporal dynamics [26]. Let D be dimensionality of the resulting exemplars and $\mathbf{A} = [\mathbf{A_s}\ \mathbf{A_n}]$ be the dictionary of size $D \times (J+K)$ used for the decomposition.

To convert a noisy recording to the exemplar space, the data is first converted to the desired time-frequency representation used to create the dictionaries. A sliding window of length T frames is moved along its time axis at a hop size of 1 frame resulting in a total of $W = L - T + 1$ windows, where L is the number of frames in the time-frequency representation. The frames corresponding to each window are reshaped to a vector and are stacked as columns in the observation data matrix $\mathbf{\Psi}$ of size $D \times W$. This is then approximated as a weighted sum of the atoms in the speech and noise dictionaries to obtain the activations $\mathbf{X}$ (of size $(J+K) \times W$) as:

$$\mathbf{\Psi} \approx \begin{bmatrix}\mathbf{A_s} & \mathbf{A_n}\end{bmatrix} \begin{bmatrix}\mathbf{X_s} \\ \mathbf{X_n}\end{bmatrix} = \mathbf{AX} \quad s.t. \quad \mathbf{X} \geq 0 \tag{1}$$

where, $\mathbf{X_s}$ and $\mathbf{X_n}$ are the activations for the speech and noise dictionaries respectively and $\mathbf{X} = [\mathbf{X_s}^\intercal\ \mathbf{X_n}^\intercal]^\intercal$. Here, $^\intercal$ denotes the matrix transpose. The approximation is done to obtain the activations $\mathbf{X}$ that minimize the generalized Kullback-Leibler divergence between $\mathbf{\Psi}$ and $\mathbf{AX}$ with additional sparsity constraint on $\mathbf{X}$, which in matrix form is formulated as:

$$\sum_{d=1}^{D}\sum_{w=1}^{W}\left\{\mathbf{\Psi}_{d,w}\log\frac{\mathbf{\Psi}_{d,w}}{(\mathbf{AX})_{d,w}} - \mathbf{\Psi}_{d,w} + (\mathbf{AX})_{d,w}\right\} + \sum_{n=1}^{(J+K)}\sum_{w=1}^{W}(\mathbf{\Lambda}\odot\mathbf{X})_{n,w} \tag{2}$$

where $\mathbf{\Lambda}$ is a matrix of size $(J+K) \times W$ which, in effect, penalizes the l_1-norm of the activations and serves as a parameter to control the sparsity of $\mathbf{X}$. $\odot$ denotes element-wise multiplication. For the rest of the paper, the subscripts $\mathbf{s}$ and $\mathbf{n}$ denote the speech and noise, respectively and the superscripts denote the exemplar spaces.

Notice that the sparsity penalty matrix $\mathbf{\Lambda}$ has a size equal to the number of atoms in the dictionary times the number of observation vectors. This matrix thus can be used to individually adjust the relative weight of any atom in the dictionary to approximate any column in the observation matrix $\mathbf{\Psi}$. However, in practise, the penalty is kept constant as λ_s for all speech atoms and λ_n for all noise atoms across all columns in the observation matrix, reducing the number of parameters to be tuned to two. $\mathbf{\Lambda}$ will thus have a structure comprised of an upper-block matrix of size $J \times W$ with all elements equal to λ_s and a lower block matrix of size $K \times W$ with all elements set as λ_n.

The cost function (2) is minimized by iteratively applying the NMF multiplicative-update rule [13,27]:

$$\mathbf{X} \leftarrow \mathbf{X} \odot \frac{\mathbf{A}^\intercal\left(\frac{\mathbf{\Psi}}{\mathbf{AX}}\right)}{\mathbf{A}^\intercal\mathbf{1} + \mathbf{\Lambda}} \tag{3}$$

where all divisions are element-wise and $\mathbf{1}$ is a matrix of ones of size $D \times W$. This update rule is the bottleneck to the processing speed and computational complexity is linear in D, J, K and W.

Once this decomposition is obtained, we can obtain the windowed estimates of speech and noise as $\hat{\mathbf{s}}_\mathbf{w} = \mathbf{A_s}\mathbf{X_s}$ and $\hat{\mathbf{n}}_\mathbf{w} = \mathbf{A_n}\mathbf{X_n}$ respectively, each of size $D \times W$. Notice that there are multiple approximations of the same time-frequency frame appearing over multiple overlapping windows of these windowed estimates. To remove this windowing effect and to obtain the frame level estimates, we first append a zero matrix of size $D \times (T-1)$ to the windowed estimate, to get a matrix of size $D \times L$, and consider it as a block matrix having T block rows of size $(D/T) \times L$ each. Let $\hat{\mathbf{s}}_{\mathbf{w},\tau}$ be the τ-th block matrix. The frame-level estimate of size $(D/T) \times L$ then obtained, similar to an overlap-add method, as:

$$\hat{\mathbf{s}} = \sum_{\tau=1}^{T} \overset{\rightarrow(\tau-1)}{\hat{\mathbf{s}}_{\mathbf{w},\tau}} \tag{4}$$

where, $\overset{\rightarrow(\tau)}{(\cdot)}$ denotes right shifting a matrix by τ columns (prepending τ columns of zeros on the left and deleting τ columns on the right so as to maintain the original matrix size during addition). Averaging by the number of overlapping windows is omitted as it will be cancelled in the later processing stages. The frame-level noise estimate $\hat{\mathbf{n}}$ is obtained in the same manner.

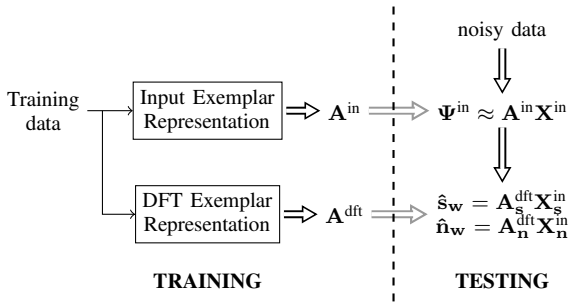


Fig. 1. Block digram overview of the proposed system using modulation spectrogram features and coupled dictionaries.

B. Method using Coupled Dictionaries

The proposed approach to obtain the DFT estimates using coupled dictionaries is summarized in Fig. 1. In this approach, the NMF-based decomposition is obtained in an additive and non-negative feature space of choice which serves as the front-end of the speech enhancement system. For simplicity, the front-end features are referred to as "input exemplars" and the dictionary used to obtain the NMF compositional model is denoted as $\mathbf{A}^{\text{in}} = \left[\mathbf{A}_{\mathbf{s}}^{\text{in}}\ \mathbf{A}_{\mathbf{n}}^{\text{in}}\right]$. This dictionary has a size $D^{\text{in}} \times (J+K)$, where D^{in} is the dimensionality of the input exemplar space. The observation data matrix in the input exemplar domain $\mathbf{\Psi}^{\text{in}}$ is decomposed using $\mathbf{A}^{\text{in}}$ as explained in section II-A. The resulting activations $\mathbf{X}^{\text{in}}$ are then applied with the output DFT dictionary to directly obtain the windowed speech and noise estimates in the DFT domain as $\hat{\mathbf{s}}_{\mathbf{w}} = \mathbf{A}_{\mathbf{s}}^{\text{dft}}\mathbf{X}_{\mathbf{s}}^{\text{in}}$ and $\hat{\mathbf{n}}_{\mathbf{w}} = \mathbf{A}_{\mathbf{n}}^{\text{dft}}\mathbf{X}_{\mathbf{n}}^{\text{in}}$, respectively.

To obtain a reliable reconstruction of the underlying DFT estimates, the mapping between the corresponding atoms in both the dictionaries should nearly be one-to-one. Such an approximation would work if the input and the output DFT exemplars are temporally aligned and scale alike with signal strength. Regarding the last criterion, signal representations that vary linearly with the input signal strength work best in conjunction with the considered cost function (2). These are achieved by properly choosing the input exemplars and extracting the corresponding DFT exemplars from the same piece of training data spanning T frames (ref. Fig. 1).

From the windowed estimates, the frame level estimates $\hat{\mathbf{s}}$ and $\hat{\mathbf{n}}$ are obtained by removing the windowing effect and the corresponding time-varying filter is obtained by element-wise division as:

$$\mathbf{W} = \hat{\mathbf{s}} \oslash (\hat{\mathbf{s}} + \hat{\mathbf{n}})\,. \tag{5}$$

This is then multiplied element-wise to the short-time Fourier transform (STFT) of the noisy speech $\mathbf{Y}$ of size $F \times L$, where F is the number of frequency bins used to obtain the STFT. The enhanced STFT, $\hat{\mathbf{S}} = \mathbf{Y} \odot \mathbf{W}$, is converted to time-domain using overlap-add method to obtain the enhanced speech. Notice that the DFT dictionary is of size $D^{\text{dft}} \times (J+K)$, where $D^{\text{dft}} = F \cdot T$ and the time-varying filter has the same size as $\mathbf{Y}$. In short, the proposed method thus can exploit the speech and noise separation capabilities for various choices of input spaces and can generate a filter which has full-rank in the DFT space.

III. Choice of input representation

The various choices for input representation that are investigated in this work are explained in this section. Notice that the underlying assumption in the exemplar-based approach is that the speech and noise are approximately additive in the chosen exemplar spaces. The processing chains for obtaining the coupled exemplars are summarized in Fig. 2.

A. DFT Exemplars

First, the DFT space is chosen as the input exemplar space to obtain the decomposition. To obtain DFT exemplars to create the DFT dictionary, a segment of length T frames (T_t seconds in time domain) of training data is chosen at random and its magnitude STFT is used for non-negativity. Let the STFT be obtained using a window length and hop size of t_w^{dft} and t_h^{dft}, respectively. This yields a spectro-temporal representation of size $F \times T$, where F is the number of frequency bins used to obtain the STFT. This is reshaped to a vector of size $(F \cdot T) \times 1$ to obtain the DFT exemplar. i.e., $D^{\text{dft}} = F \cdot T$.

During evaluation, the NMF-based decomposition is done in the DFT space after converting the noisy observation into its equivalent DFT exemplar representation. The resulting activations are used to obtain the frame-level speech and noise estimates, and the enhanced speech is obtained as explained in Section II. This setting is chosen as one of the baseline systems in this work and is denoted as *DFT-DFT* setting.

B. Mel Exemplars

Mel exemplars are chosen for their lower dimensionality and robust speech and noise separation performance in the presence of a variety of noises. First, the Mel features for T frames of data are obtained after applying Mel-integration of the magnitude STFT as depicted in Fig. 2. This is done by multiplying the magnitude STFT by the DFT-to-Mel matrix $\mathbf{M}$ which contains the magnitude response of B Mel bands along its rows. The resulting representation of size $B \times T$ is reshaped to a vector to obtain the Mel exemplar of length $D^{\text{mel}} = B \cdot T$. The Mel dictionaries for speech and noise are denoted as $\mathbf{A}_{\mathbf{s}}^{\text{mel}}$ and $\mathbf{A}_{\mathbf{n}}^{\text{mel}}$, respectively.

During the test phase, the noisy data represented in the Mel exemplar space is decomposed using the Mel dictionary $\mathbf{A}^{\text{mel}} = [\mathbf{A}_{\mathbf{s}}^{\text{mel}}\ \mathbf{A}_{\mathbf{n}}^{\text{mel}}]$ and the corresponding activations $\mathbf{X}_{\mathbf{s}}^{\text{mel}}$ and $\mathbf{X}_{\mathbf{n}}^{\text{mel}}$ are obtained. Once these activations are obtained, we use it to evaluate two systems.

First, another baseline system is defined which is denoted as the *Mel-Mel* setting. In this setup, the windowed speech and noise estimates are obtained using the Mel dictionary as $\mathbf{A}_{\mathbf{s}}^{\text{mel}}\mathbf{X}_{\mathbf{s}}^{\text{mel}}$ and $\mathbf{A}_{\mathbf{n}}^{\text{mel}}\mathbf{X}_{\mathbf{n}}^{\text{mel}}$, respectively. The frame level Mel estimates, $\hat{\mathbf{s}}'$ and $\hat{\mathbf{n}}'$ are obtained as explained in Section II-A. These are then mapped to the DFT domain using the pseudo-inverse of the DFT-to-Mel matrix, $\mathbf{M}^{\dagger} = \mathbf{M}^{\intercal}(\mathbf{M}\mathbf{M}^{\intercal})^{-1}$ to obtain the enhanced STFT as [12]:

$$\hat{\mathbf{S}} = \mathbf{Y} \odot \left(\mathbf{M}^{\dagger}\left[\hat{\mathbf{s}}' \oslash (\hat{\mathbf{s}}' + \hat{\mathbf{n}}')\right]\right). \tag{6}$$

It is evident that this setting has a lower computational complexity as $B \ll F$ while performing the multiplicative updates. It is also observed that Mel features have a better speech and noise separation capability and generalize better for unseen noise cases when compared to the DFT exemplars [21]. However, the pseudo-inverse mapping in (6) will always fall in a subspace of rank B spanned by the rows of $\mathbf{M}$. The frequency response of the Mel filter-bank being triangular, such a mapping is equivalent to a piece-wise linear approximation of B points located at the central frequencies of the filter-bank. It is thus clear that such a transformation may not be able to model most of the speech and noise content in the full-resolution DFT space with $B \ll F$, which in turn may reduce the speech enhancement quality. This issue will be further explored in later sections.

For the second setting, we investigate the proposed approach using Mel exemplars as the input features to deal with the low-rank approximation in the Mel-Mel setting. Here, the underlying (windowed) DFT estimates for speech and noise are directly obtained

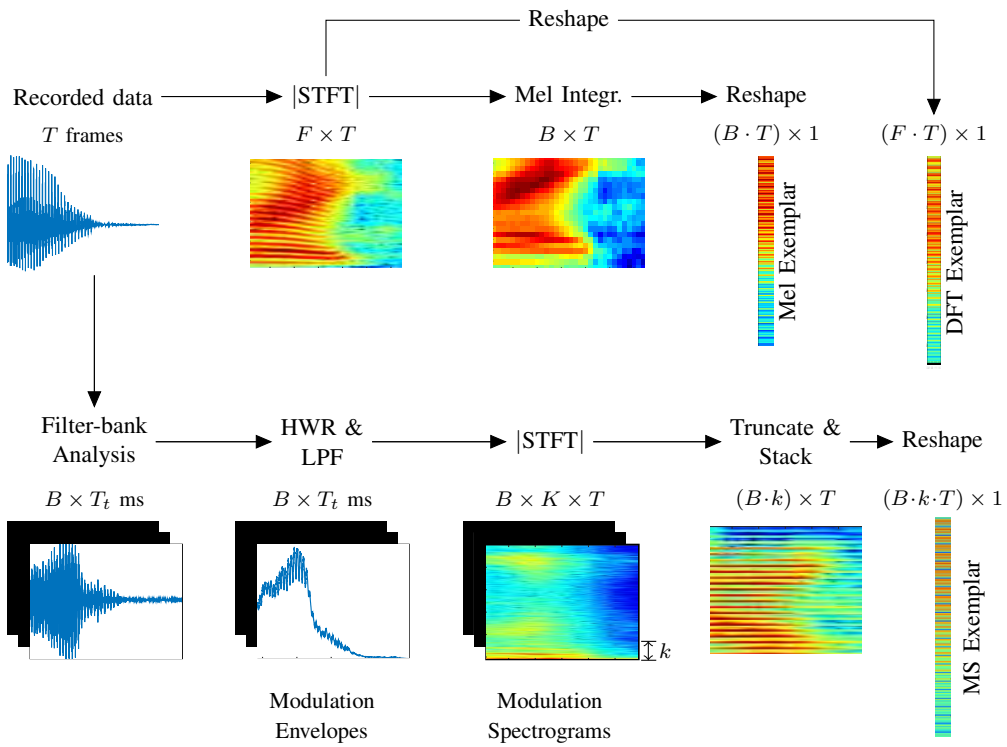


Fig. 2. Block diagram overview of the processing chains to obtain various exemplars. All the coupled exemplars are extracted from the same piece of recorded data spanning T frames (T_t seconds in time-domain). The resulting representations along with their size are shown below in each of the steps. Figures are not shown at the same scale.

using the Mel activations as $\hat{\mathbf{s}}_\mathbf{w} = \mathbf{A}_\mathbf{s}^{\text{dft}}\mathbf{X}_\mathbf{s}^{\text{mel}}$ and $\hat{\mathbf{n}}_\mathbf{w} = \mathbf{A}_\mathbf{n}^{\text{dft}}\mathbf{X}_\mathbf{n}^{\text{mel}}$, and are then used for speech enhancement (ref. Section II). This is referred to as the *Mel-DFT* setting. Since in this setting, the output DFT dictionary is coupled to the Mel input dictionary and is overcomplete, a full-rank reconstruction of the estimates can be enforced and a better noise suppression could be achieved.

C. MS Exemplars

The modulation spectrogram (MS) representation of speech was proposed as part of a computational model for human hearing which relies on low frequency amplitude modulation variations within frequency bands [28]. These variations play a key role in the higher level human auditory processing [29] and are computationally modelled as modulation envelopes. The bottom row in Fig. 2 summarizes the processing chain to obtain the MS representation for speech.

To obtain the modulation envelopes, the acoustic data is first filtered using a filter bank containing B channels to model the frequency discrimination property of the basilar membrane. The resulting B bandlimited signals are half-wave rectified to model the non-negative nerve firings followed by low-pass filtering to obtain the modulation envelopes. The 3dB cut-off frequency of the low-pass filter used is around 20Hz as human speech contains modulations of very low frequency [30] and hence the spectrograms of these envelopes, called the modulation spectrograms, can yield a more effective representation [25]. Let the window length and hop size used to obtain the MS representation be t_w^{MS} and t_h^{MS}, respectively. The MS representation is typically obtained over longer window lengths when compared to the DFT features (i.e., $t_w^{\text{MS}} > t_w^{\text{dft}}$), to capture the variation in modulation envelopes, which also allows larger choices for t_h^{MS} than t_h^{dft}. This representation of speech has successfully been used for blind source separation [31] and noise-robust ASR [32].

Notice that converting acoustic data into the MS space results in a three-dimensional representation of size $B \times K \times T$, where B, K and T are the number of input frequency channels, number of modulation frequency bins and number frames in the acoustic data, respectively. However, since the modulation envelopes are obtained after a low-pass filtering operation, only a few bins in the MS will contain significant energy and it is possible to truncate each of the MS to the lowest few, say k, bins. These truncated B modulation spectrograms, each of size $k \times T$, are stacked to get a two-dimensional representation of size $(B \cdot k) \times T$, referred to as the *MS features*. This representation is then reshaped to a vector to obtain the MS exemplar. The dimensionality of an MS exemplar will thus be $D^{\text{MS}} = B \cdot k \cdot T$. In our previous works [21,33] we showed that the approximate additivity assumption of speech and noise is valid in the MS exemplar space as well. In comparison to the established Mel exemplar-based approaches, the MS representation essentially retains the same information within each frequency band for each frame, but also more accurate information about the spectral distribution of different modulation frequencies.

In this work, the MS exemplars are used as input exemplars to obtain the NMF-based decomposition using the dictionary of MS exemplars $\mathbf{A}^{\text{MS}} = [\mathbf{A}_\mathbf{s}^{\text{MS}}\ \mathbf{A}_\mathbf{n}^{\text{MS}}]$ to obtain the activations $\mathbf{X}^{\text{MS}}$. However, since the processing chain to obtain the MS features involves non-linear operations, there is no direct way to make use of this decomposition to enhance the noisy speech as the inversion of the MS features to the time domain is not unique. We propose using the coupled DFT dictionary extracted together with the MS dictionary to reconstruct the DFT estimates and to obtain speech enhancement, i.e., the speech and noise estimates are approximated as $\mathbf{A}_\mathbf{s}^{\text{dft}}\mathbf{X}_\mathbf{s}^{\text{MS}}$ and $\mathbf{A}_\mathbf{n}^{\text{dft}}\mathbf{X}_\mathbf{n}^{\text{MS}}$, respectively. The resulting frame-level estimates are used to enhance the noisy spectrogram. This system is denoted as the *MS-DFT* setting.

However, any circular temporal shift (modulo the window length) of the DFT spectrogram can yield the same MS representation and makes the mapping many-to-one. To address this, we make use of temporal oversampling, i.e., smaller t_h^{MS} while obtaining the MS, to reduce this ambiguity as pointed out in [34]. In our previous work, setting $t_h^{\text{MS}} = t_h^{\text{dft}}$ was found to be the best choice [33]. It is also to be noted that increasing the low-pass cut of frequency beyond 20 Hz should be useful for a better speech and noise separation when the data is corrupted by some noise having higher modulation frequencies. This on the other hand requires a higher value of k which increases the computational complexity and may lead to data overfitting. Hence, a compromise must be pursued.

IV. Experimental Setup

A. Databases

To evaluate and compare the various settings, two databases were used. Preliminary experiments were conducted on the AURORA-2 database which is a small-vocabulary task and are then extended to the large-vocabulary database AURORA-4 [35].

1) AURORA-2 Database: is a database based on the TI Digits corpus containing utterances of digits from '0-9' and 'oh' sampled at 8 kHz. For training the acoustic models, a clean speech dataset and a noisy training dataset each containing $8\ 440$ utterances are used. The noisy training set contains car, babble, subway and exhibition hall noises added artificially at signal-to-noise ratios (SNRs) of 5, 10, 15 and 20 dB.

For testing, test sets A and B are used. Test set A contains one clean subset containing $1\ 001$ recordings of clean speech and its noisy versions at varying SNRs -5, 0, 5, 10, 15 and 20 dB for every noise type present in the training set, summing to a total of 28 subsets. Test set B also has the same structure as in test set A but with four different noise types which are not present in the training data. The noise types in test set B are restaurant, train station, street and airport noises.

2) AURORA-4 Database: is a large vocabulary continuous speech database based on the WSJ-0 corpus of read speech. The database contains training and test sets with additive noise and in presence of channel variation. In this work, only the single microphone test set with 16 kHz sampling frequency, which contains a noise free data set (test 01 or test A) with six noisy sets (test 02-07 or collectively test B) corrupted with car, babble, restaurant, airport, street and train noises added artificially at varying SNRs between 5 and 15 dB in steps of 1 dB, is used. A development set of the same structure as the test set, but with different utterances, is also there for validation and parameter tuning.

For training the acoustic models and preparing the dictionaries, the clean and the multi-noise training sets containing $7\ 138$ utterances each were used. The multi-noise training set contains all noises present in the test sets added at varying SNRs between 10 and 20 dB in steps of 1 dB.

B. Exemplars and dictionary preparation

The dictionaries used to obtain the decomposition were prepared from the training data. The noise data used to create the noise exemplars were obtained from the noisy training data using the two-step procedure described in [27]. The dictionaries were created using exemplars originating from random segments of length T frames taken from the clean and noise training sets. Throughout this paper, the choice of T used was 30 and 15 frames for the AURORA-2 and AURORA-4 databases, respectively as these values were found to yield the best performance on similar tasks [10,12]. The value of T for AURORA-4 is chosen to be lesser than the AURORA-2 database as the former has a lot more variety of speech to be modelled as opposed to the latter and it demands a larger dictionary to reasonably model the large vocabulary speech data, which increases the computational complexity.

Every chosen random segment of length T_t seconds was first pre-processed by removing the DC component and applying a pre-emphasis filter (a single order high-pass filter of coefficient 0.97). The coupled exemplars were then extracted as follows (ref. Fig. 2):

1) The STFT of the samples were obtained using a Hamming window of length $t_w^{\text{dft}} = 25$ ms and a hop size $t_h^{\text{dft}} = 10$ ms. The magnitude of the STFT is then obtained yielding a representation of size $F \times T$. This is then reshaped to obtain the DFT exemplar of length $F \cdot T$.
2) The magnitude STFT obtained in the step above is pre-multiplied with the DFT-to-Mel matrix $\mathbf{M}$ of size $B \times F$ to obtain the Mel-integrated magnitude spectra of size $B \times T$. The Mel exemplar is then obtained by reshaping the Mel spectra.
3) To obtain the MS representation, the time-domain signal is first filtered into B band-limited signals using the equivalent rectangular bandwidth filter banks implemented using Slaney's toolbox [36]. Each of these signals is then half-wave rectified and low-pass filtered at a 3 dB cut-off frequency of 30 Hz (as used in [33]) to obtain the modulation envelopes. The MS representation is then obtained by taking the magnitude STFT of these envelopes by keeping the hop size $t_h^{\text{MS}} = t_h^{\text{dft}} = 10$ ms and using a window length $t_w^{\text{MS}} = 64$ ms as in [33]. $K = 64\text{ms} \times f_s$ frequency bins are used to obtain the STFT, where f_s is the sampling frequency. i.e., the first frequency bin corresponds to ≈ 15 Hz resulting in approximately 3 frequency bins below 30 Hz cut-off frequency including the DC component. A value of $k = 5$ is chosen to capture the frequency leakage during low-pass filtering and windowing. The MS exemplar is then obtained as detailed in Section III-C. Notice that the number of channels in the filter bank is the same as the number of Mel filters used in the previous step. This choice is made to have a fair comparison between the performances of the Mel and the MS exemplars in separating speech and noise.

For the experiments on the AURORA-2 database, the parameters used were $F = 128$ and $B = 23$ used were whereas the AURORA-4 setting used were $F = 256$ and $B = 40$. Zero-padding was used while taking the STFT, whenever necessary. Then three coupled dictionaries each for speech and noise were created with the corresponding exemplars extracted from the same piece of training data.

To create the speech dictionary, $J = 10\ 000$ exemplars were extracted at random from the respective clean training data for experiments on the AURORA-2 and AURORA-4 databases as used in [21,33]. Evaluations on the AURORA-2 database used a noise dictionary containing $K = 10\ 000$ exemplars, whilst for the AURORA-4 experiments, the noise dictionary used is comprised of two parts: a fixed noise dictionary containing $K_{\text{fixed}} = 5\ 000$ exemplars extracted from the noise training data and a small noise dictionary extracted from the noisy test data to be enhanced itself, which are the cyclicly shifted versions of its first $T = 15$ frames resulting in a total of $K = 5\ 015$ noise exemplars as in [12,33]. Making use of the first 15 frames to model the noise is termed as *noise-sniffing* assuming the first 15 frames of the noisy test data contain noise only. Notice that the second noise dictionary is changed for every utterance and is concatenated with the fixed noise dictionary.

Extracting the fixed part of the coupled dictionaries was done only once per database and are kept fixed for all the experiments

in this paper. The noise dictionaries for the AURORA-2 database contain exemplars sampled from all the four noise types available in the training data and the fixed noise dictionary for AURORA-4 experiments contain all the six noise types in the training data. No supervision was done to avoid silences in the speech exemplars or adjusting the number of exemplars per noise type in the noise dictionary.

C. NMF based speech enhancement

For testing, the noisy utterance is converted to the input exemplar space to obtain the observation data matrix $\mathbf{\Psi}^{\text{in}}$ as explained in Section II. $\mathbf{\Psi}^{\text{in}}$ is then decomposed using the respective input dictionary using 600 NMF multiplicative updates (3) with $\mathbf{X}^{\text{in}}$ initialised as $\left(\mathbf{A}^{\text{in}}\right)^{\intercal}\mathbf{\Psi}^{\text{in}}$ and the corresponding filters are obtained as described in Section III. The resulting enhanced STFT is inverted to the time-domain using the overlap-add method to obtain the enhanced speech.

For the AURORA-2 setting, the decomposition was obtained with speech and noise sparsity penalties as $\lambda_s = 1.5$ and $\lambda_n = 1$ for the Mel dictionary as used in [27] whilst for the decomposition using the MS and DFT dictionaries, the values used were $\lambda_s = 1.75$ and $\lambda_n = 0.75$ as in [21]. These values were obtained after doing a grid-search in the range $[0, 3]$ on a development set which is a subset of 100 files taken from the test set A.

For the AURORA-4 experiments, in contrast to the AURORA-2 setting, the noise sparsity penalty is fixed as 0.5 times the sparsity penalty of speech, i.e., $\lambda_n = \lambda_s/2$, to reduce the computational effort while doing the grid-search [12] on the development set. The decomposition using the Mel, MS and DFT settings used a λ_s equal to 1.2, 1.6 and 1.7 respectively.

Speech enhancement was implemented using MATLAB and GPUs were used for accelerating the NMF multiplicative updates using the parallel computing toolbox. To evaluate and compare the speech enhancement qualities, we used signal-to-distortion ratio (SDR), segmental SNR (SegSNR) and PESQ measurements. SDRs were obtained using the BSS evaluation toolkit [37], and the other two measurements were calculated using an implementation by Loizou [38]. The improvement of these quality measures over the noisy speech is reported as ΔSDR in dB, ΔPESQ in mean opinion score (MOS) and ΔSegSNR in dB..

D. ASR back-ends

1) HMM-GMM decoder for AURORA-2: For evaluating the ASR performance on the AURORA-2 database, a GMM-HMM-based recognizer using the Mel-frequency cepstral coefficients (MFCCs) was used. The HMM topology had a total of 179 states comprised of 16 states describing each digit with 3 states for silence $(16 \times 11 + 3)$. GMM models were trained on MFCCs with 13 static coefficients along with the delta and delta-delta coefficients leading to a 39 dimensional feature space. The emission probabilities of each of the HMM states were modelled using a GMM of 32 Gaussians with diagonal covariance. The decoding is done using the Viterbi decoder with a finite state language model as given in the AURORA-2 benchmark [35] with all digits having the same word entrance penalties.

2) Hybrid setting for AURORA-2: Preliminary experiments on the AURORA-2 database revealed a complementarity in the number of insertions and deletions between the MS-DFT and the Mel-DFT systems. So a hybrid approach is proposed to combine the outcomes of these two recognizers to achieve a better ASR performance. There exist several ways to combine results from two systems like assuming independence and then balance the two streams [10], minimum error based approach [39], etc. In order to avoid extra parameters, we propose to combine the two streams by simply multiplying the likelihoods originating from the Mel-DFT and MS-DFT settings [10]. Equal weights are given to both the streams by raising both the resulting likelihoods by 0.5. i.e.,

$$p'(y_t|q_t) = \left(p_{\text{mel}}(y_t|q_t)\right)^{1/2} \left(p_{\text{MS}}(y_t|q_t)\right)^{1/2} \tag{7}$$

where, p_{mel} and p_{MS} are the likelihoods for the observation y_t given the HMM state q_t resulting from the Mel-DFT and MS-DFT streams, respectively. These are then fed to the Viterbi decoder to obtain the ASR results.

3) HMM-GMM decoder for AURORA-4: For the AURORA-4 experiments, the "recipe" recognizers in the Kaldi toolkit [40] are used. The HMM-GMM-based recipe decoder for AURORA-4 makes use of context dependent tied-state triphone models. Each model is comprised of three states and there are around 2000 distinct HMM states in total. GMM models are trained on 13 static MFCC features from 7 consecutive frames upon which feature decorrelation is applied using maximum-likelihood linear transform (MLLT) [41] and linear discriminant analysis (LDA) [42], reducing the 91-dimensional vector to 40 dimensions.

4) DNN-HMM decoder for AURORA-4: In this work, we also evaluate the ASR performance using the DNN-HMM hybrid system, where the posterior probability estimates for the HMM states are provided by the trained DNNs [43]. DNNs are comprised of multiple hidden layers stacked on top of each other which allow them to learn higher-level information in the upper layers [44]. The recipe recognizer is based on the implementation described in [45] with 6 hidden layers comprised of 2048 sigmoid neurons per layer. The input layer used 40 Mel filterbank coefficients with a context size of 11 frames summing up to 440 input features in total.

To train the DNN, pre-training based on restricted Boltzmann machines (RBMs) [46] is done first in order to avoid issues with random initialization of the layers resulting in poor local optima. Once the pre-training is done, a DNN which classifies the frames into triphone states is trained using the stochastic gradient descent technique. Finally, the DNN is trained to classify the whole sentence correctly. For the DNNs trained on clean training data, only the clean part of the development set was used for cross-validation.

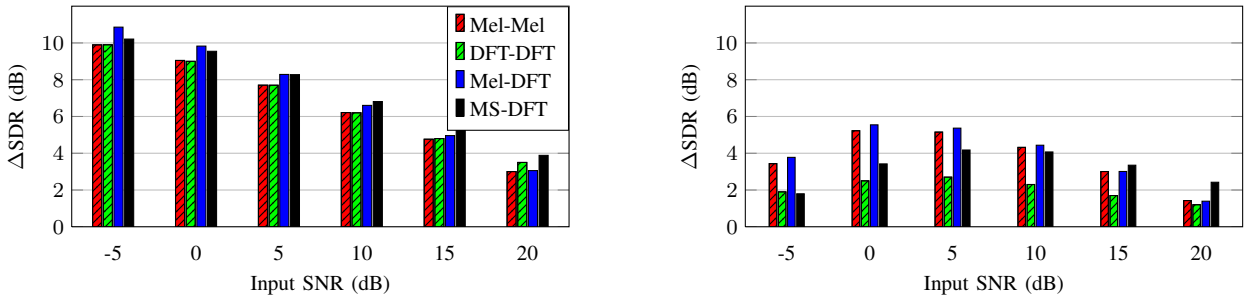
Average word error rates (WERs) are used as the performance measure in all the ASR experiments. For training the acoustic models, the original clean training data (referred to as *clean training*) and the enhanced noisy training data processed by the corresponding NMF-based front-ends (referred to as *retraining*) are used. Retraining equips the GMMs and DNNs to learn the artefacts introduced by the enhancement stage and thus can improve the ASR performance on the enhanced noisy test data.

V. Pilot Experiments on AURORA-2

This section details the speech enhancement and ASR evaluations performed on the AURORA-2 database. The results are reported on the entire test sets including the 100 files used for tuning the sparsity parameters. Some useful insights and discussions are also included in this section.

A. Results on speech enhancement

ΔSDR in dB averaged over the four noise types obtained for various systems on the AURORA-2 database are summarized in Fig. 3. The shaded bars denote the baseline systems and it can be seen that the proposed approach using coupled dictionaries results in better SDRs in all cases. Notice that, even though the Mel-Mel setting uses a pseudo-inverse, it yields almost the same SDRs as of the DFT-DFT setting on test set A. This can be attributed to the better



(a) Test set A (b) Test set B

Fig. 3. Average SDR improvements in dB obtained on test sets A and B of the AURORA-2 database as a function of input SNRs in dB for various settings. Legends are same for both plots.

TABLE I
AVERAGE WERS IN % OBTAINED FOR TEST SETS A AND B OF THE AURORA-2 DATABASE FOR VARIOUS SETTINGS WITH GMMS TRAINED ON CLEAN AND ENHANCED NOISY TRAINING DATA. SHADED ROWS DENOTE THE BASELINE SETTINGS. BEST SCORES ARE HIGHLIGHTED IN BOLD FONT.

Setting	clean	Test Set A							Test Set B						
		-5	0	5	10	15	20	Avg. (20-0)	-5	0	5	10	15	20	Avg. (20-0)
GMM on clean training data															
No Enhancement	0.3	76.9	48.7	22.4	9.2	3.6	1.6	17.1	77.2	46.9	20.7	7.7	2.8	1.2	15.9
Mel-Mel	0.4	31.2	12.4	6.1	3.6	2.3	1.4	5.2	58.2	30.3	**12.4**	5.8	2.7	0.9	10.4
DFT-DFT	**0.3**	34.7	17.5	7.8	3.1	1.7	0.9	6.2	70.8	40.1	16.9	6.1	2.3	1.0	13.3
Mel-DFT	0.4	31.1	12.4	6.0	3.5	2.1	1.2	5.0	**58.0**	**30.1**	**12.4**	5.7	2.7	0.8	**10.3**
MS-DFT	**0.3**	30.5	12.5	4.4	**2.1**	**1.3**	**0.7**	4.2	68.6	34.3	14.5	**5.1**	2.1	0.8	11.4
Hybrid	0.4	**27.2**	**11.4**	**3.7**	**2.1**	1.5	0.9	**3.9**	62.4	32.8	13.0	5.3	**2.0**	**0.6**	10.7
GMM on noisy training data (Retrained)															
No Enhancement	0.8	61.9	24.9	6.8	2.6	1.2	0.7	7.2	64.3	26.2	8.5	2.9	1.4	0.8	8.0
Mel-Mel	0.5	25.1	8.9	3.3	1.5	0.9	1.0	3.1	52.8	20.8	6.8	2.6	1.2	0.7	6.4
DFT-DFT	**0.4**	21.4	8.5	2.5	1.1	**0.7**	0.5	2.7	58.1	24.5	7.5	2.4	**1.0**	0.6	7.2
Mel-DFT	0.5	25.2	9.0	3.1	1.4	**0.7**	0.6	3.0	**52.6**	21.0	6.8	2.7	1.2	0.6	6.4
MS-DFT	**0.4**	21.1	7.7	**2.4**	1.0	**0.7**	**0.4**	**2.4**	62.4	26.3	7.6	2.2	**1.0**	**0.5**	7.5
Hybrid	**0.4**	**20.6**	**7.1**	**2.4**	**1.0**	**0.7**	0.7	**2.4**	54.2	**20.7**	**6.3**	**2.1**	**1.0**	**0.5**	**6.1**

speech and noise separation achieved by the Mel exemplars when compared to the DFT exemplars. It can also be seen that the Mel-DFT setting yields better SDRs than the Mel-Mel setting for both test sets, even though the decomposition in both the systems are done in the Mel exemplar space. It reveals the effectiveness of using the proposed coupled DFT dictionary approach to directly obtain the DFT estimates over the low-rank approximation using pseudo-inverse.

From the SDR evaluations on test set B which contains unseen noise cases, it can be seen that the speech enhancement obtained is poorer when compared to that of test set A, as the noise dictionary generalises poorly to the unseen noise cases. It can also be seen that the Mel feature space is able to better generalise to the unseen noise cases when compared to the DFT and MS exemplar spaces. Using the proposed Mel-DFT approach can further increase the SDR performance, which is a scenario where the proposed approach is highly beneficial. It can also be seen that the MS space can yield a better speech and noise separation at high SNRs when compared to the Mel features.

B. ASR evaluation

The average WERs obtained on the enhanced AURORA-2 data using the HMM-GMM based decoder and also using the hybrid setting described in Section IV-D are summarized in Table I for GMMs trained on the clean training data (clean training) and the enhanced noisy training data (retrained). It can be seen that the method using coupled dictionaries yields improved WERs and retraining the GMMs using the enhanced training data can further improve the ASR performance. The Mel-DFT setting resulted only in a slight improvement when compared to the Mel-Mel setting, even though the former setting yielded a better speech enhancement in terms of SDRs. This can be attributed to the simplicity of the AURORA-2 recognition task as it has a limited vocabulary, and the digit classification is not affected by the deformation introduced during the pseudo-inverse step.

It is also observed that the use of the MS representation can result in a WER improvement for test set A and poorer results for test set B as it generalises poorly for unseen noise cases. Nevertheless, it yielded complementary results in terms of insertions and deletions when compared to the Mel setting and the proposed hybrid setting was found to yield superior WER improvements on both test sets by exploiting this complementarity. To the best of our knowledge, average WERs of 20.6% (test A, SNR-5), 2.4% (test A, SNR(20-0)) and 6.1% (test B, SNR(20-0)) using the hybrid setting are among the best results ever reported on the AURORA-2 recognition task (reported in [10]). Overall, from SNR -5 dB to 20 dB, the hybrid setting yielded WERs of 5.4% and 14.1% on test set A and B, respectively.

Also notice that the method described in [27] directly makes use of enhanced Mel features for the ASR back-end rather than going back to the time-domain. Evaluations (not shown) revealed that this setting and the Mel-Mel setting are equivalent as the ASR back-end for the latter also goes back to the Mel domain by multiplication with the same Mel matrix $\mathbf{M}$ to obtain the MFCCs.

C. A qualitative analysis

A qualitative analysis on the observations made during the pilot experiments on the AURORA-2 database is discussed in this section. The outcomes of interest resulting from these evaluations are visualised in Fig. 4. The input noisy signal is an arbitrary signal from

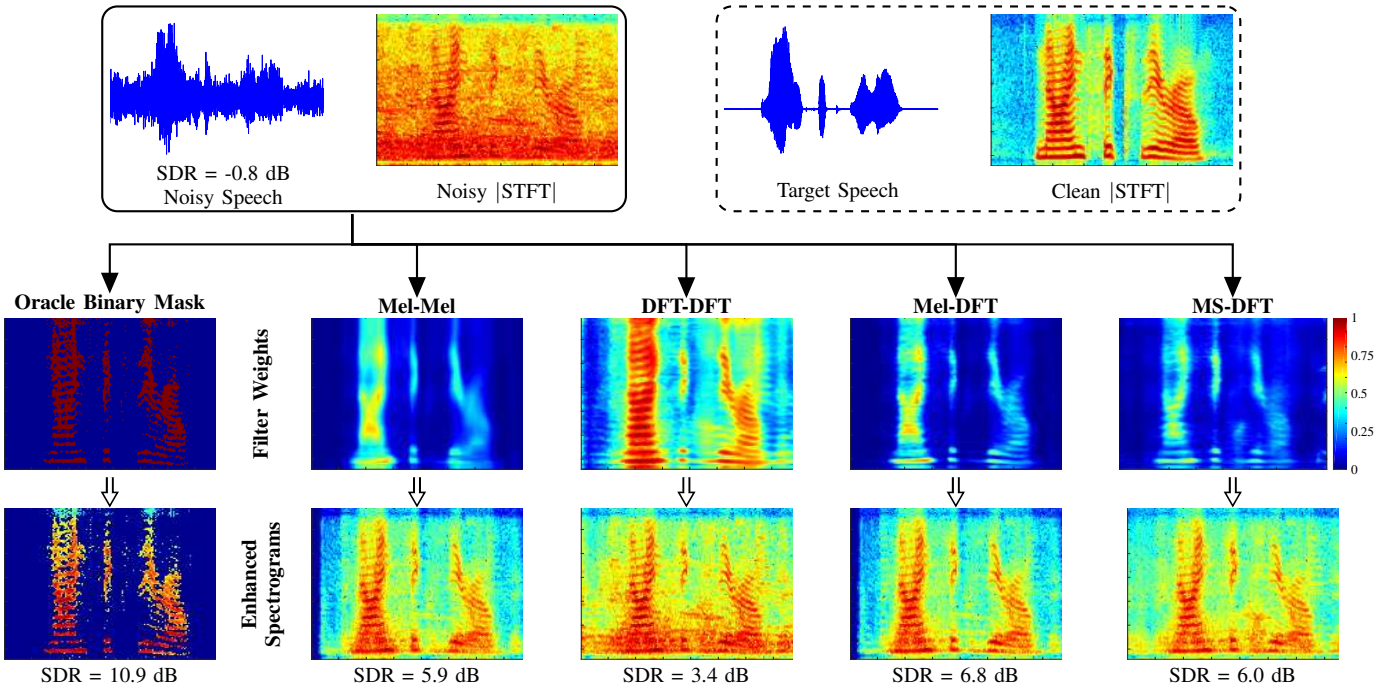


Fig. 4. Comparison of filter weights with oracle binary mask and the resulting enhanced speech spectrograms obtained for various settings for an arbitrary noisy speech signal from the AURORA-2 database corrupted with babble noise at an SNR of 0 dB. Log spectrograms are shown for a better visualisation. All filter weight plots used a linear color mapping in $[0, 1]$. The resulting SDRs are also shown below each of the enhanced spectrograms.

TABLE II
AVERAGE EXECUTION TIME IN SECONDS NEEDED FOR VARIOUS SETTINGS EVALUATED ON THE AURORA-2 DATABASE. D IS THE NUMBER OF ROWS IN THE DICTIONARY USED TO OBTAIN THE NMF-BASED DECOMPOSITION. ALL DICTIONARIES HAD A TOTAL OF $20,000$ COLUMNS EACH.

	Mel-Mel	*DFT-DFT*	*Mel-DFT*	*MS-DFT*
Exec.time	5.8s	16.2s	6.0s	14.8s
D	690	3840	690	3450

the AURORA-2 database containing the utterance "nine six zero" (transcribed as *96Z*) corrupted with babble noise at an SNR of 0 dB. The filter weights used for enhancing the noisy STFT arising from the various settings are shown in the middle row followed by the resulting enhanced speech in the bottom row. For comparison, the oracle binary mask is also included which yielded an output SDR of 10.9 dB. It is evident that the quality of enhanced speech depends on how well these filter weights model the constituent speech and noise contained in the noisy speech. The key aspects which decide the performance of various settings are detailed below (ref. Fig. 4):

1) Low-rank approximation in the Mel-Mel setting: It can be seen that the piece-wise linear approximation results in a set of filter weights that are smooth which in turn cannot model the underlying harmonic structure of the constituent speech signal and results in frequency smearing. This setting thus will always result in a sub-optimal set of filter weights. Also notice that this setting still yielded a reasonable SDR improvement which can be attributed to a better speech and noise separation achieved using the Mel exemplars.

2) Poorer speech and noise separation in the DFT exemplar space: It can be seen that the filter weights arising from the DFT-DFT setting are able to model the underlying harmonic structure of speech since this setting can directly obtain the estimates in the full-resolution frequency domain. However, a majority of these weights are close to 1 even though the true SNR of the underlying speech is 0 dB, which in turn retain most of the noise content and results in poorer SDRs. Also notice that the noise in the speech inactive regions are not properly suppressed. These happen because the speech exemplars are also activated to model the babble noise contained in the noisy input during the exemplar-based decomposition in the DFT space. Similar instances of speech exemplars modelling noise are observed for unseen noise cases also (not shown) [21]. This setting hence results in a poorer SDR improvement even though the detrimental mapping stage is absent.

3) Full-rank approximation in the Mel-DFT setting: The filter weights obtained for the Mel-Mel and Mel-DFT settings arise from the same set of activations obtained from the NMF-based decomposition in the Mel exemplar space. It can be seen that the Mel-DFT approach is able to better model the harmonic structure in speech and utilise the better speech and noise separation properties of the Mel exemplar space, yielding an SDR improvement of 0.9 dB over the setting where the pseudo-inverse is used. This approach thus can yield a better speech enhancement without any additional computational cost in the matrix factorisation part, which is the most time-consuming part of the method.

4) Coupled dictionaries as a reliable mapping from the MS space to the DFT/time domain: It is evident from the filter weights obtained for the MS-DFT setting that the MS exemplars can yield a good speech and noise separation, and using the coupled DFT dictionary can yield a reliable mapping of these estimates to the full-resolution frequency domain.

D. Computational complexity vs performance

All the evaluated experiments in this work were accelerated using GPUs. The computational complexity of these experiments depends on the length of the temporal context T, the number of exemplars $(J + K)$ and the dimension of features per frame considered. The average execution time needed for the experiments on the AURORA-2 database, which used $10\ 000$ exemplars each of speech and noise with $T = 30$ frames for various settings are tabulated in Table II.

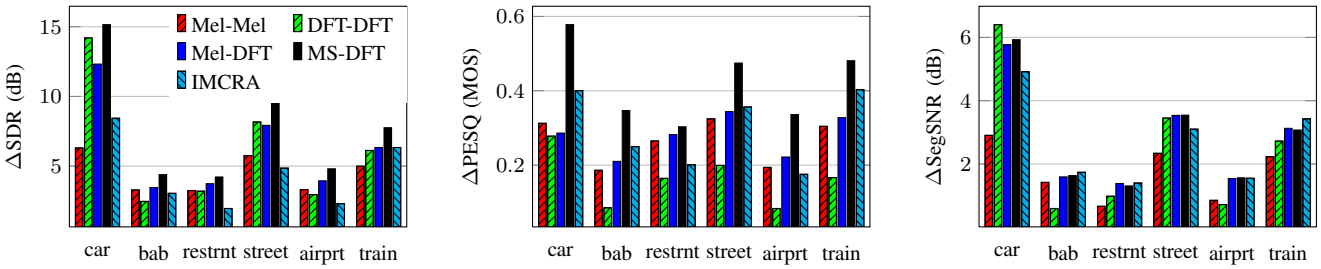


Fig. 5. Average improvements in speech enhancement performance in terms of ΔSDR, ΔPESQ and ΔSegSNR obtained for each test set on the AURORA-4 database for various settings. From left to right, these noises correspond to test02-07 (car, babble, restaurant, street airport and train noises, respectively). The legends are same for all plots.

TABLE III
AVERAGE WERS OBTAINED IN % FOR VARIOUS TEST SETS ON THE AURORA-4 DATA USING THE VARIOUS SETTINGS WITH THE HMM-GMM-BASED AND HMM-DNN-BASED ASR BACK-ENDS. BEST SCORES ARE HIGHLIGHTED IN BOLD FONT. SHADED ROWS DENOTE THE BASELINE SYSTEMS.

(a) Retrained GMM

Setting	Test Sets							
	A	B						
	01	02	03	04	05	06	07	Avg.
No Enh.	5.7	6.2	11.5	22.3	16.7	10.9	15.8	13.9
Mel-Mel	5.1	5.6	8.4	10.6	9.8	8.1	10.1	8.8
DFT-DFT	6.0	5.8	8.9	12.2	10.3	8.7	11.2	9.5
Mel-DFT	4.9	**5.4**	8.0	**10.7**	9.8	7.7	10.2	8.6
MS-DFT	4.9	5.7	**7.3**	11.1	**9.0**	**7.0**	**10.1**	**8.4**
IMCRA	**4.6**	5.6	10.7	15.3	13.8	11.4	14.4	11.9

(b) Retrained DNN

Setting	Test Sets							
	A	B						
	01	02	03	04	05	06	07	Avg.
No Enh.	3.3	4.6	7.3	9.3	8.5	6.6	9.1	7.7
Mel-Mel	**2.9**	**4.1**	6.6	8.8	8.9	6.1	9.2	7.3
DFT-DFT	3.2	**4.1**	6.9	7.8	7.6	6.5	8.0	6.8
Mel-DFT	3.2	4.7	7.5	8.5	8.4	6.9	8.2	7.4
MS-DFT	3.0	4.2	**6.0**	**7.4**	**7.1**	**5.3**	**6.9**	**6.2**
IMCRA	**2.9**	**4.1**	7.2	9.4	9.6	7.6	9.0	7.8

From the evaluations, it is clear that the proposed Mel-DFT setting results in a good ASR and SDR performance without much additional computational cost.

It is also observed in [33] that increasing the low-pass 3 dB cut-off frequency in the MS exemplar extraction stage can yield an improvement both in terms of SDRs and WERs, in presence of seen noise cases. However, this can have a detrimental effect for signals corrupted with unseen noise and also results in an increased computational complexity as the size of the MS exemplars should also be increased.

Similar to the MS features, the performance of the DFT exemplars depend on the type of noise and the true SNRs in the input noisy signal, and its computational complexity depends solely on the sampling frequency of the input data, given T and the window length t_w^{dft} used to obtain the STFT. On the other hand, the Mel and MS features are more flexible in the sense that their dimensionality can be adjusted by varying choices for B, t_w^{MS} etc., depending on the application and allowable computational complexity.

VI. EXPERIMENTS ON AURORA-4 DATABASE

A. Results on speech enhancement

ΔSDR, ΔPESQ and ΔSegSNR averaged per test set obtained for the various settings on the AURORA-4 database are presented in Fig. 5. As an additional baseline system, a speech enhancement algorithm based on minimum mean-square error log-spectral amplitude estimation [47] with the improved minima controlled recursive averaging (IMCRA) technique for noise variance estimation [48] is included.

It can be seen that the proposed approach using coupled dictionaries results in better SDRs in all cases, consistent with the observations made during the AURORA-2 experiments. It can also be seen that additional evaluations using the PESQ and SegSNR also yielded promising improvements. IMCRA approach yielded better SegSNR for some noise types, but poorer PESQ and SDR improvements were obtained. The MS-DFT setting yielded superior improvements in PESQ MOS evaluation reaffirming the effectiveness of using coupled dictionaries to obtain a reliable reconstruction in the DFT space.

B. ASR evaluation

The average WERs obtained for the HMM-GMM-based and HMM-DNN-based decoders on various test sets of the NMF-enhanced AURORA-4 data are tabulated in Table III. The results for the retrained scenarios only are presented for both the GMM and DNN based settings.

For acoustic modelling based on retrained GMMs, it can be seen that the various speech enhancement approaches can greatly improve the ASR performance over a GMM trained and evaluated on noisy test data. IMCRA yields the best performance on clean speech as it introduces the least distortions on clean speech during speech enhancement. It can also be seen that the MS-DFT setting yields the best performance out of all the evaluated settings with a statistical significance of $p < 0.03$ (over a total of 32 118 words using a binomial independence assumption).

On the other hand, a DNN trained on noisy training data yields around 40% relative improvement over the GMM-based system and is even better than the best performing retrained GMM setting (ref. Table IIIa), thanks to its multiple hidden layers which can learn and compensate for the noise also. It can be seen that using exemplar-based approaches for speech enhancement and retraining can further improve its performance (ref. Table IIIb). Also notice that all settings yielded a better WER for clean speech as well, which can be attributed to the ability of sparse representations in moving the test features closer to the training features, thereby minimizing the speaker mismatches in the training and test sets as pointed out in [49].

The MS-DFT setting yielded the best WERs here as well with a statistical significance of $p < 0.001$ over all the other settings yielding an average WER of 6.2% over test B of the AURORA-4 database.

VII. Conclusions

In this work, we proposed using coupled DFT dictionaries, extracted jointly with the input dictionaries used in the exemplar-based speech enhancement systems, for a better mapping from the input space to the DFT space to obtain a better set of filter weights. The approach was found to be effective in overcoming the low-rank approximation where the input dictionary is created using lower-dimensional Mel features and also to obtain a reliable mapping from the MS space to the DFT space. The simulation results revealed that the proposed approach can improve the performance of exemplar-based techniques for both speech enhancement and automatic speech recognition tasks.

The use of modulation spectrogram features, which are inspired from the human auditory processing, was also introduced to the field of exemplar-based techniques in this work, and we showed that using coupled dictionaries can be a reliable way to reconstruct the underlying speech and noise estimates in the DFT domain. The ASR evaluation also revealed that feeding NMF-enhanced data can greatly benefit both the HMM-GMM-based and DNN-HMM-based state-of-the-art ASR systems with and without retraining.

The best performing settings in this work yielded overall average WERs of 5.4% and 14.1% respectively for test sets A and B of the AURORA-2 database, and 7.9% and 5.7% respectively for the GMM-HMM-based and DNN-HMM-based ASR systems on the single microphone sets (test01-test07) in the AURORA-4 database.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, April 1979.

[2] V. Grancharov, J. Samuelsson, and B. Kleijn, "On causal algorithms for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 764–773, 2006.

[3] J. Jensen, J. Benesty, M. Christensen, and S. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 1948–1963, 2012.

[4] T. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *Speech and Audio Processing, IEEE Transactions on*, vol. 4, no. 5, pp. 383–389, 1996.

[5] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *Signal Processing, IEEE Transactions on*, vol. 40, no. 4, pp. 725–735, 1992.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, October 1999.

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

[8] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 1–12, Jan 2007.

[9] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *International Conference on Spoken Language Processing*, 2006.

[10] J. F. Gemmeke and H. Van hamme, "Advances in noise robust digit recognition using hybrid exemplar based systems," in *INTERSPEECH*. ISCA, 2012.

[11] A. Hurmalainen and T. Virtanen, "Modelling spectro-temporal dynamics in factorisation-based noise-robust automatic speech recognition," in *Acoustics, Speech and Signal Processing, 2012 IEEE International Conference on*, 2012, pp. 4113–4116.

[12] J. T. Geiger, J. F. Gemmeke, B. Schuller, and G. Rigoll, "Investigating NMF speech enhancement for neural network based acoustic models," in *INTERSPEECH*. ISCA, 2014.

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

[14] M. N. Schmidt and J. Larsen, "Reduction of non-stationary noise using a non-negative latent variable decomposition," in *IEEE Internatioanal Workshop on Machine Learning and Signal Processing XVIII*. IEEE, October 2008.

[15] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear mmse filter for single channel speech enhancement based on nonnegative matrix factorization," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, Oct 2011, pp. 45–48.

[16] J. Le Roux, J. R. Hershey, and F. Weninger, "Sparse NMF – half-baked or well done?" in *Techical Report*, no. TR2015-023. Cambridge, MA, USA: Mitsubishi Electric Research Labs (MERL), Mar. 2015.

[17] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2009, pp. 1705–1713.

[18] N. Mohammadiha and S. Doclo, "Single-channel dynamic exemplar-based speech enhancement," in *INTERSPEECH*. ISCA, 2014.

[19] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *INTERSPEECH*. Makuhari, Japan: ISCA, 2010.

[20] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2140–2151, 2013.

[21] D. Baby, T. Virtanen, T. Barker, and H. Van hamme, "Coupled dictionary training for exemplar-based speech enhancement," in *Acoustics, Speech and Signal Processing, 2014 IEEE International Conference on*, May 2014, pp. 2883–2887.

[22] J. Nam, G. J. Mysore, J. Ganseman, K. Lee, and J. S. Abel, "A super-resolution spectrogram using coupled PLCA," in *INTERSPEECH*, Makuhari, Japan, 2010, pp. 1696–1699.

[23] Z. Wu, T. Virtanen, T. Kinnunen, E. Chng, and H. Li, "Exemplar-based unit selection for voice conversion utilizing temporal information." in *INTERSPEECH*. ISCA, 2013, pp. 3057–3061.

[24] M. Gönen, "Coupled dimensionality reduction and classification for supervised and semi-supervised multilabel learning," *Pattern Recognition Letters*, vol. 38, pp. 132–141, 2014.

[25] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Acoustics, Speech, and Signal Processing, 1997 IEEE International Conference on*, vol. 3, 1997, pp. 1647–1650.

[26] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Acoustics, Speech and Signal Processing, 2010 IEEE International Conference on*, March 2010, pp. 4546 –4549.

[27] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067 –2080, Sept. 2011.

[28] C. Plack, *The sense of hearing*. Lawrence Erlbaum Associates Publishers, 2005.

[29] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. The MIT Press, 1990.

[30] C. E. Schreiner and J. V. U., "Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF)," *Hearing Research*, vol. 21, pp. 227 – 241, 1986.

[31] T. Barker and T. Virtanen, "Non-negative tensor factorization of modulation spectrograms for monaural sound source separation," in *INTERSPEECH*. ISCA, 2013.

[32] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram." *Speech Communication*, vol. 25, no. 1-3, pp. 117–132, 1998.

[33] D. Baby, T. Virtanen, J. F. Gemmeke, T. Barker, and H. Van hamme, "Exemplar-based noise robust speech recognition using modulation spectrogram features," in *Spoken Language Technology Workshop, 2014 IEEE*, South Lake Tahoe, USA, December 2014.

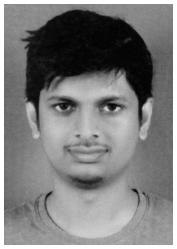
[34] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," in *Acoustics, Speech, and Signal Processing, 1983 IEEE International Conference on*, vol. 8, 1983, pp. 804–807.

[35] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, 2000, pp. 29–32.

[36] M. Slaney, "Auditory toolbox version 2," *Interval Research Corporation*, vol. 10, 1998.

[37] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.

[38] P. C. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, 1st ed. CRC Press, 2007.
[39] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Automatic Speech Recognition and Understanding, 1997 IEEE Workshop on*, 1997, pp. 347–354.
[40] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.
[41] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Acoustics, Speech, and Signal Processing, 2000 IEEE International Conference on*, vol. 2, 2000, pp. 1129–1132.
[42] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1992 IEEE International Conference on*, vol. 1, Mar 1992, pp. 13–16 vol.1.
[43] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.
[44] M. A. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2014.
[45] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *INTERSPEECH*. ISCA, 2013, pp. 2345–2349.
[46] G. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural Networks: Tricks of the Trade (2nd ed.)*, 2012, pp. 599–619.
[47] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, Apr 1985.
[48] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 466–475, Sept 2003.
[49] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representation features for speech recognition," in *INTERSPEECH*. ISCA, 2010, pp. 2254–2257.



Deepak Baby received the Bachelors degree in Electronics and Communication Engineering from College of Engineering, Trivandrum, India in 2009 and Masters degree in Communication and Signal Processing from Indian Institute of Technology, Bombay in 2012. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering, KU Leuven, Belgium.

His research interests are noise-robust automatic speech recognition, machine learning, speech enhancement and compressed sensing.

Tuomas Virtanen Tuomas Virtanen is an Academy Research Fellow and Associate Professor (tenure track) at Department of Signal Processing, Tampere University of Technology (TUT), Finland, where he is leading the Audio Research Group. He received the M.Sc. and Doctor of Science degrees in information technology from TUT in 2001 and 2006, respectively. He has also been working as a research associate at Cambridge University Engineering Department, UK.

He is known for his pioneering work on single-channel sound source separation using non-negative matrix factorization based techniques, and their application to noise-robust speech recognition, music content analysis and audio event detection. In addition to the above topics, his research interests include content analysis of audio signals in general and machine learning. He has authored more than 100 scientific publications on the above topics, which have been cited more than 3000 times. He has received the IEEE Signal Processing Society 2012 best paper award for his article "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria" as well as three other best paper awards. He is an IEEE Senior Member and recipient of the ERC 2014 Starting Grant.

Jort Florent Gemmeke (*jgemmeke@amadana.nl*) is a research scientist at Google. Prior to this, he worked as algorithm developer at Audience, Mountain View. In 2011, he received the Ph.D degree from the University of Nijmegen, The Netherlands, on the subject of noise robust ASR using missing data techniques. From 2011 to 2014, he worked as a postdoc at KU Leuven, Belgium, working on a project to develop self-taught vocal interfaces for dysarthric speakers.

He is known for pioneering the field of exemplar-based noise robust ASR employing sparse non-negative matrix factorization, for which he was awarded the 2014 IEEE young author best paper award. His research interests are acoustic event detection, automatic speech recognition and source separation.

Hugo Van hamme received the Masters degree in engineering (burgerlijk ingenieur) from VUB in 1987, the M.Sc. degree from Imperial College, London in 1988 and the Ph.D. degree in electrical engineering from Vrije Universiteit Brussel (VUB) in 1992. From 1993 till 2002, he worked for L&H Speech Products and ScanSoft, initially as senior researcher and later as research manager. Since 2002, he is a professor at the department of electrical engineering of KU Leuven.

His main research interests are: applications of speech technology in education and speech therapy, computational models for speech recognition and language acquisition and noise robust speech recognition.