Speech Enhancement Using Generative Dictionary Learning

Christian D. Sigg, *Member, IEEE*, Tomas Dikk, and Joachim M. Buhmann, *Senior Member, IEEE*

***Abstract*—The enhancement of speech degraded by real-world interferers is a highly relevant and difficult task. Its importance arises from the multitude of practical applications, whereas the difficulty is due to the fact that interferers are often nonstationary and potentially similar to speech. The goal of monaural speech enhancement is to separate a single mixture into its underlying clean speech and interferer components. This under-determined problem is solved by incorporating prior knowledge in the form of learned speech and interferer *dictionaries*. The clean speech is recovered from the degraded speech by sparse coding of the mixture in a *composite dictionary* consisting of the concatenation of a speech and an interferer dictionary. Enhancement performance is measured using *objective measures* and is limited by two effects. A too sparse coding of the mixture causes the speech component to be explained with too few speech dictionary atoms, which induces an approximation error we denote *source distortion*. However, a too dense coding of the mixture results in *source confusion*, where parts of the speech component are explained by interferer dictionary atoms and vice-versa. Our method enables the control of the source distortion and source confusion trade-off, and therefore achieves superior performance compared to powerful approaches like *geometric spectral subtraction* and *codebook-based filtering*, for a number of challenging interferer classes such as speech babble and wind noise.**

***Index Terms*—Dictionary learning, sparse coding, speech enhancement.**

I. Introduction

ENHANCING speech degraded by nonstationary real-world interferers is both an important and difficult task. The importance arises from many signal processing applications, including hearing aids, mobile communications, and preprocessing for speech recognition. The difficulty of speech enhancement in these applications arises from the nature of the encountered interferers, which often are nonstationary and potentially speech-like, thereby inducing a significant and time-varying spectral overlap between speech and interferer.

Manuscript received April 27, 2011; revised October 08, 2011; accepted December 31, 2011. February 06, 2012; date of current version March 30, 2012. This work was supported in part by CTI grant 8539.2;2 ESPP-ES. Date of publication The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sharon Gannot.

C. D. Sigg is with the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss), Zurich, Switzerland (e-mail: christian@sigg-iten.ch).

T. Dikk and J. M. Buhmann are with the Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland (e-mail: tomasdikk@tomasdikk.com; jbuhmann@inf.ethz.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2187194

The goal of speech enhancement is twofold: to improve both the perceived *quality* and the *intelligibility* of speech, by attenuating the interferer without substantially degrading the speech. Speech of higher quality is perceived as being more comfortable to listen to, for longer periods of time, whereas higher speech intelligibility is measured by lower word error rates in speech recognition scenarios.

Ideally, the performance of speech enhancement algorithms is measured by conducting subjective listening tests with human listeners. *Objective measures* are designed to approximate subjective quality scores and intelligibility rates. Most objective measures quantify improvement by comparing the (unobserved) clean speech with the degraded speech and the enhanced speech in a perceptually meaningful way. As a consequence, performance evaluation has to be conducted on synthetic mixtures of clean speech and interferer signals.

We consider the setting of a one-to-one conversation in a natural environment, recorded by a single microphone. This setup can be modeled as a linear additive mixture of *target* clean speech and *interferer*

$$x(n) = s(n) + i(n) \tag{1}$$

where $x(n)$ is the time-domain mixture signal at sample n, and $s(n)$ and $i(n)$ are the time-domain speech and interferer signals. Recovering the clean speech signal from the mixture is underdetermined without additional assumptions. Our enhancement approach is based on transforming time-domain signals into a suitably chosen feature space, and *sparse coding* in this feature space using signal models for both the speech and the interferer (called *dictionaries*). Since speech and many kinds of interferers contain structure, their structured component can be sparsely coded in *coherent* dictionaries. If both the speech and interferer dictionary is coherent only to its respective structured component in the mixture signal, sparse coding is able to separate the mixture into its structured components and to suppress any unstructured component (i.e., random noise) that is *incoherent* to both dictionaries. Finally, an estimate of $s(n)$ is obtained by performing the inverse transform from the feature space back to the time-domain.

Since clean speech is never observable in the environment where enhancement is to take place, we learn the speech dictionary on a training corpus. Speech is a well-structured signal class, therefore a pre-trained model remains largely valid during enhancement, even in the speaker independent case. The contrary is true for the interferer, which varies considerably depending on the environment, and which might be a superposition of several sources, requiring a single general interferer

model to be prohibitively complex. On the other hand, the interferer can be observed during segments of speech inactivity. Therefore, training data for an interferer model can be obtained from speech pauses, resulting in an interferer dictionary which is specific to the current environment and which does not have to generalize to other environments. In this work, we presuppose that a conservative voice activity detector (VAD) is available to obtain observations of the interferer signal for dictionary learning. A VAD which is not conservative enough could cause speech signal components to be present in the interferer dictionary training data, and as a consequence, enhancement performance likely degrades. Evaluation using a non-ideal VAD is beyond the scope of this paper and is considered in future work. For the speech enhancement itself, no knowledge of speech activity is necessary. This paper extends our preliminary results reported in [34].

A. Related Work in Speech Enhancement

The various approaches to speech enhancement can be categorized based on the assumptions made about the speech and the interferer signals, and how these assumptions are exploited for estimating the clean speech from the mixture.

Our method falls into the class of environment-adapted algorithms, which incorporate specific knowledge about the environment where speech enhancement is to be performed. In this paper, we focus on speech enhancement methods which infer an environment-specific interferer model. Spectral subtraction (Section I-A1) employs a point estimate of the average interferer magnitude spectrum. Codebook-based spectral filtering (Section I-A2) approximates the distribution of both the speech and interferer magnitude spectra using a codebook of spectral prototypes. Finally, source separation based on sparse coding (Section I-A3) models both the speech and the interferer signal classes using a sparse linear combination of atoms from a respective dictionary.

Environment-adapted methods based on different principles include SNR classification of time-frequency bins for binary masking [17] and optimizing the sub-band weighting rule for a specific environment and perceptual cost function [15].

1) Spectral Subtraction: Spectral subtraction is historically one of the first enhancement algorithms [3]. The transformation of the mixture (1) into the short-time Fourier domain,

$$X(\omega, n) = S(\omega, n) + I(\omega, n) \tag{2}$$

where $X(\omega, n)$, $S(\omega, n)$ and $I(\omega, n)$ denote the complex mixture, speech and interferer spectra at frequency ω and time n, respectively, suggests the following principle: an estimate of the clean speech spectrum is obtained by subtracting an estimate of the interferer spectrum from the observed mixture spectrum. Typically, only the magnitude $|I(\omega, n)|$ of the spectrum or the power spectrum $|I(\omega, n)|^2$ is estimated (e.g., during speech pauses), and as a consequence the additivity of (2) only holds in approximation. The time-domain speech signal estimate is obtained by inverse Fourier transformation of the estimated speech magnitude spectrum or power spectrum using the phase of the mixture signal.

In many relevant cases, modeling the interferer signal using a single spectral prototype is insufficient, due to the nonstationarity of the encountered interferers. As a consequence, estimation errors can lead to negative values in the estimated clean speech spectrum. Basic spectral subtraction algorithms set these negative values to zero (or some floor value). This nonlinear processing causes isolated peaks in the spectrum, which is one cause for distracting residual *musical noise* in the time-domain estimate of clean speech.

Geometric spectral subtraction [21] preserves the additivity of (2) by taking the phase of the complex Fourier coefficients into account. In contrast to spectral subtraction, the estimate of the interferer spectrum is updated for each frame using minimum statistics [26]. This method is further discussed in Section V-A

2) Codebook-Based Spectral Filtering: Whereas spectral subtraction employs a point estimate of the interferer spectrum, codebook-based spectral filtering models either the speech, the interferer or both using vector codebooks. Ellis and Weiss [13] trained a speaker dependent codebook using vector quantization (VQ), and projected the mixture onto the closest clean speech prototype for enhancement. Srinivasan *et al.* [36] trained linear prediction coefficient (LPC) codebooks for both the speech and the interferer. The observed mixture spectrum is assumed to be a linear combination of exactly two spectral prototypes (with associated gains), one from the speech and the other from the interferer codebook. The selected prototype pair is used to estimate the underlying clean speech by Wiener filtering of the mixture signal. Roweis [30] enforces a temporal dependency between consecutive codings of observations using a factorial hidden Markov model.

Although codebooks are much more sophisticated signal models, the major drawback of this paradigm is the induced quantization error due to the maximally sparse coding. To reduce the quantization error to an acceptable level, a very large speech codebook [13], interpolation [36] or averaging over all possible vector pairs [37] is required (see Section V-B for further discussion).

3) Sparse Coding for Source Separation: Structured signal classes like speech have approximately sparse representations in suitably chosen dictionaries. This key observation underlies source separation methods based on sparse coding. In *coherent denoising* [25], an estimate of clean speech can be recovered from a mixture of speech and Gaussian white noise by capturing the components that are coherent to the speech dictionary, because unstructured interferers such as Gaussian white noise are incoherent to any fixed dictionary [28]. A dictionary element (called *atom*) is coherent to a signal if the absolute value of the inner product of the two vectors is large. For an orthonormal basis as the speech dictionary, the energy in the coding coefficients of an unstructured interferer is distributed uniformly over all dictionary elements. A soft thresholding of the coding coefficients results in a close to optimal estimate of the clean speech [9].

However, as interferers become more structured and similar to speech, the coherence to the speech dictionary grows. In the case of structured interferers, better performance can be achieved by coding the mixture in a composite dictionary

consisting of a speech *and* an interferer dictionary. Speech components are captured by the speech dictionary, as they are more coherent to the speech dictionary, but less so to the interferer dictionary. For the same reason, structured interferer components are captured by the interferer dictionary instead of being explained by many speech dictionary atoms with low associated weights. Any unstructured interferer component is again incoherent to both dictionaries.

Therefore, sparse coding for speech enhancement requires adapted dictionaries, as generic analytic dictionaries (such as wavelet bases) typically do not satisfy both coherence and incoherence requirements. Adaptation of an initial dictionary to the signal class can be achieved using dictionary learning, which is a generalization of VQ [1]. From this perspective, the codebook-based enhancement approaches of the previous section can be seen as a special case of sparse coding for speech enhancement, where a maximally sparse coding of the mixture is enforced.

There exists an analogy between sparse coding in learned dictionaries and subspace methods for enhancement. Subspace methods estimate the speech signal subspace using the singular value decomposition (SVD) and recover the clean speech from an orthogonal projection of the mixture signal onto the speech signal subspace [14]. The assumption of orthogonality between the speech and interferer subspaces is relaxed in the case of enhancement based on sparse coding in learned dictionaries. For a sparse enough coding of the underlying speech and interferer signals, a nonzero mutual coherence between the speech and interferer dictionaries is tolerated (Section IV-B), which corresponds to nonorthogonal speech and interferer signal subspaces.

B. Dictionary Learning

Dictionary learning performs approximate matrix factorization of a data matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$ into the product of a dictionary matrix $\mathbf{D} \in \mathbb{R}^{D \times L}$ and a coding matrix $\mathbf{C} \in \mathbb{R}^{L \times N}$, under some sparsity constraints on the coding matrix. Dictionary learning is the generalization of gain-shape codebook learning. Signal vectors are represented as linear combinations of multiple dictionary atoms, allowing for lower approximation error while maintaining equal dictionary size.

Our dictionary learning method is based on the K-SVD algorithm [1]. K-SVD is an iterative method alternating between sparse coding and dictionary update steps. At iteration t, the training data $\mathbf{X}$ is sparsely coded in dictionary $\mathbf{D}^{(t-1)}$ to obtain coding $\mathbf{C}^{(t)}$, followed by a dictionary update to obtain $\mathbf{D}^{(t)} = \arg\min_{\mathbf{D}} ||\mathbf{X} - \mathbf{D}\mathbf{C}^{(t)}||_F^2$, which minimizes the approximation error given the current coding $\mathbf{C}^{(t)}$ (see Section I-F for a definition of our notation).

In the special case of nonnegative features (such as Fourier magnitudes), nonnegative matrix factorization (NMF) [18] provides a different matrix factorization where the elements of both $\mathbf{C}$ and $\mathbf{D}$ are constrained to be nonnegative. Sparse NMF includes an additional sparsity constraint on $\mathbf{C}$ [12], [16].

C. Method Overview

What follows is a high-level overview of our method, illustrated in Fig. 1. The time-domain signal is transformed (*FT* in the figure) into either the short-time Fourier (STFT) magnitude domain or the modified discrete cosine (MDCT) domain [27]. The feature transform needs to be invertible to allow for the resynthesis of the time-domain signal of the enhanced speech. The MDCT is directly invertible, but in the case of STFT magnitudes, the mixture phase is used for resynthesis. In the transform domain, overlapping blocks are extracted and vectorized, and these vectors form the elements of our feature space (Section VI-C).

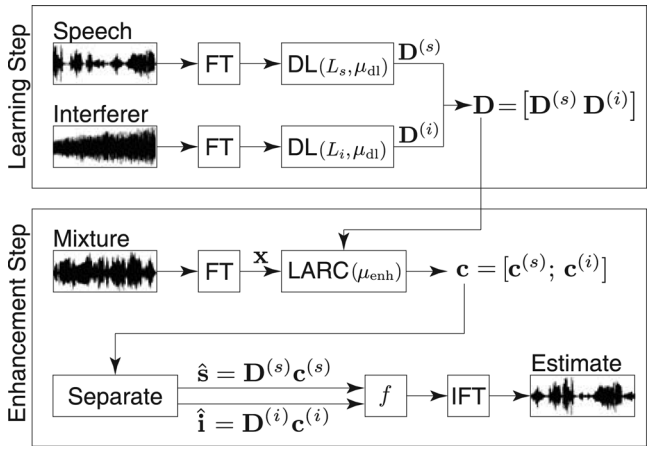


Fig. 1. Our dictionary learning and speech enhancement pipeline. The top part shows the dictionary learning step, where, separately for the speech and the environment specific interferer, training data is transformed into the feature space (*FT*), and a dictionary is learned (*DL*). Both dictionaries $\mathbf{D}^{(s)}$ and $\mathbf{D}^{(i)}$ are combined into a composite dictionary $\mathbf{D}$ for speech enhancement. The bottom part shows the enhancement step for a single observation of degraded speech. It is transformed into the feature space, and $\mathbf{x}$ is sparsely coded in $\mathbf{D}$ using LARC. The sparse code $\mathbf{c}$ is separated into $\mathbf{c}^{(s)}$ and $\mathbf{c}^{(i)}$, to obtain estimates $\hat{\mathbf{s}}$ and $\hat{\mathbf{i}}$ of the clean speech and interferer contributions to $\mathbf{x}$. For the MDCT feature domain, $\hat{\mathbf{s}}$ is directly inverse transformed (*IFT*) back to the time domain. For the STFT magnitude domain, a filter f is built based on $\hat{\mathbf{s}}$ and $\hat{\mathbf{i}}$ and applied to $\mathbf{x}$, followed by the IFT of the filtered $\mathbf{x}$ back to the time domain.

Possibly over-complete dictionaries (Section III) are trained for speech (either speaker dependent or independent) and the environment specific interferer (DL in the figure). For the sparse coding step (Section II) in the dictionary learning algorithm, we extended the *least angle regression* (LARS) algorithm [11] to include a residual coherence stopping criterion and optimized it to solve a large number of simultaneous coding problems efficiently. A C implementation of the algorithm with a Matlab interface is available from the authors.[1] For the dictionary update step, we use the fast approximate SVD update of [32]. The same algorithm parameters (such as dictionary size, residual coherence threshold or number of iterations) are used for training both the speech and the interferer dictionary, in all tested environments. The trained dictionaries are concatenated to form the *composite dictionary*.

In the enhancement step (Section IV, bottom half of Fig. 1), an observation of degraded speech is sparsely coded in the composite dictionary. As a result, the mixture of speech and interferer is explained by a linear combination of atoms from the speech dictionary and of atoms from the interferer dictionary. For the MDCT, the clean speech estimate is directly inverse transformed (*IFT* in the figure) back to the time domain. For the STFT magnitude domain, a filter is built from the clean speech

[1]http://sigg-iten.ch/research/taslp2012/

and interferer magnitude estimates, and is applied to the mixture magnitude. The filtered mixture magnitude is combined with the mixture phase to resynthesize the time domain signal.

As will be explained in Section IV-B, estimation errors result from two different and competing effects. A too sparse coding of the degraded speech in the composite dictionary induces an approximation error of the clean speech, which we call *source distortion*. A too dense coding avoids source distortion, but causes *source confusion*, by explaining some of the speech energy using interferer atoms. In order to achieve low source distortion for a sparse coding, the dictionaries must be coherent to their respective signal class. To avoid source confusion, the trained dictionaries must have low mutual coherence, i.e., the speech dictionary must be incoherent to the interferer signal, and vice-versa. Good speech enhancement performance can be achieved by choosing a feature space which is high dimensional enough such that low mutual coherence becomes feasible, and by training sufficiently powerful dictionaries such that a low approximation error is achieved with sparse codings.

D. Evaluation of Enhancement Performance

We evaluate our speech enhancement method both in the speaker independent and the speaker dependent case, using nonstationary interferer data recorded in real environments (Section VI-A). The performance of our enhancement method is quantified by the frequency-weighted segmental SNR [39], an objective measure that has been shown to correlate well with subjective judgment of speech quality and intelligibility [22]. Further objective measure results (e.g., PESQ scores) and sound clips for subjective evaluation are available from the authors' website.

We compare our approach against two baseline algorithms, which make more restrictive assumptions about the speech and/or the interferer signal: geometric spectral subtraction (Section V-A) and codebook-based spectral filtering (Section V-B). Enhancement experiments are performed using synthetic mixtures of clean speech and interferer, at A-weighted speech to interferer power ratios (SIRs) ranging from low (10-dB SIR) to high degradation (0-dB SIR). The low-degradation scenario reveals distortions introduced by the enhancement algorithms themselves, while the high degradation scenario tests the interferer attenuation capability. In each algorithm, the same parameter values are used for all interferers and all mixture SIRs.

E. Contributions of Paper

Our algorithm attains significant speech quality and intelligibility improvements (Section VI) in challenging environments, where the interferer signal is nonstationary, potentially similar to speech and where the SIR is low. At the same time, it introduces minimal distortions to the speech signal of its own.

We propose an extension of the least angle regression (LARS) algorithm of Efron *et al.* [11] for dictionary learning and sparse coding, where instead of using the residual norm or the coding cardinality as the stopping criterion, the algorithm terminates if the coherence between the current residual and the dictionary is below a chosen threshold. Furthermore, the algorithm is reformulated such that coding many observations in the same dictionary becomes much more efficient. Our algorithm is called LARC, for least angle regression with a coherence criterion.

We propose a blocking scheme in the feature domain (discussed in Section VI-C), which enables a tradeoff between time and frequency resolution of the dictionary atoms. Enhancement can be performed in the MDCT domain, which facilitates direct inversion of the speech estimate, or it can be performed in the STFT domain, where either the speech magnitude estimate is combined with the mixture phase, or the speech and interferer magnitudes provide estimates of the instantaneous *a priori* and *a posteriori* SNR, from which a suppression rule is derived for filtering the mixture magnitude (see Section V-A).

Contrary to spectral subtraction, our approach does not assume a stationary interferer, and contrary to denoising by sparse coding in analytic dictionaries, it can attenuate interferers that are partially coherent to the speech signal. Our approach is a generalization of codebook-based enhancement methods, and it achieves significantly lower source distortion because it is not restricted to a one-sparse coding of the source.

Our algorithm has conceptual similarities to the nonstationary noise reduction algorithm of Schmidt and Larsen [33], which is based on a nonnegative latent variable decomposition model of the speech and the interferer signal, and to work by Wilson *et al.* [41], which employs NMF to train speech and interferer models, as well as to work by Smaragdis *et al.* [35], which applies probabilistic latent component analysis (PLCA) to train speech and interferer models. All these approaches enforce nonnegativity constraints on both the dictionary and the coding matrices, and they are therefore only appropriate for nonnegative feature domains. Our approach allows the code and the dictionary entries to assume values of the entire real domain. Furthermore, the residual coherence stopping criterion of LARC is invariant to changes in signal energy, and the same residual coherence threshold works well across different interferer signal classes and mixture SIRs, which are not known in advance.

Notation

Given a matrix $\mathbf{A}$ and a column vector $\mathbf{b}$, scalars i, j, and *ordered* sets $\mathcal{I}$, $\mathcal{J}$ of scalars, $A_{i,j}$ denotes the scalar matrix element on row i and column j; $\mathbf{a}_{(i,:)}$ denotes the ith row-vector of matrix $\mathbf{A}$; $\mathbf{a}_{(:,j)}$ denotes the jth column-vector of matrix $\mathbf{A}$; $\mathbf{A}_{(\mathcal{I},\mathcal{J})}$ denotes the sub-matrix of $\mathbf{A}$, consisting of all elements of $\mathbf{A}$ which are on rows indexed by $\mathcal{I}$ *and* on columns indexed by $\mathcal{J}$. $\mathbf{b}^\top$ denotes a row-vector, $\mathbf{b}_\mathcal{I}$ denotes the column-vector consisting of all elements of $\mathbf{b}$ indexed by $\mathcal{I}$, and b_i denotes the ith element of vector $\mathbf{b}$. Sequences of matrices, vectors and scalars are indexed by $\mathbf{A}^{(i)}$, $\mathbf{b}^{(i)}$ and $c^{(i)}$.

II. Sparse Coding

The goal of sparse coding is to approximate a signal observation with low error, using a linear combination of only a few signal prototypes from a prespecified set of prototypes. More formally, a *K-sparse coding* $\mathbf{c} \in \mathbb{R}^L$ of a single signal frame $\mathbf{x} \in \mathbb{R}^D$ in a dictionary $\mathbf{D} \in \mathbb{R}^{D \times L}$ of unit norm atoms $\|\mathbf{d}_{(:,l)}\|_2 = 1,\ \forall l = 1, \ldots, L$ defines a sparse linear combination of $K \ll L$ atoms. The *cardinality* $\|\mathbf{c}\|_0 = K$ is the

number of nonzero coefficients of $\mathbf{c}$, and is also called the ℓ_0 pseudo-norm of $\mathbf{c}$. The dictionary $\mathbf{D}$ can be *over-complete*, i.e., $L > D$ is possible (and often desired).

Sparse coding lies at the core of our method, both in dictionary learning and in enhancement. It is a tradeoff between three factors: the signal approximation error $\|\mathbf{x} - \mathbf{D}\mathbf{c}\|_2$ measured using the ℓ_2 norm, the coding cardinality and the dictionary size. For structured signal classes such as speech or music, a dictionary exists such that a low approximation error can be achieved with a sparse coding. Such a dictionary is said to be *coherent* to the signal class, and the approximation error decays quickly as the coding cardinality increases. On the other hand, white noise is an unstructured signal class that is *incoherent* to any fixed dictionary [28], and the error decays slowly as the cardinality increases. Of course, zero approximation error is possible in both cases if $\mathbf{D}$ spans the signal space, but for a coherent dictionary the error will be already sufficiently small for $K \ll D$.

The sparse coding problem can be formulated using a cardinality constraint

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \|\mathbf{x} - \mathbf{D}\mathbf{c}\|_2 \quad \text{s.t. } \|\mathbf{c}\|_0 \leq K \tag{3}$$

or using an error constraint

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_0 \quad \text{s.t. } \|\mathbf{x} - \mathbf{D}\mathbf{c}\|_2 \leq \sigma. \tag{4}$$

Solving either (3) or (4) using the ℓ_0 pseudo-norm is an NP-hard combinatorial problem [6], an approximation scheme is therefore necessary. We discuss greedy optimization in Section II-A and convex relaxation of the ℓ_0 norm to the ℓ_1 norm in Section II-B.

A. Orthogonal Matching Pursuit

Orthogonal matching pursuit (OMP) [7] computes an approximate solution to the sparse coding problem (3) or (4) using a greedy iterative update of $\mathbf{c}$ (see algorithm 1).

Algorithm 1 Orthogonal Matching Pursuit

1: Input: $\mathbf{x} \in \mathbb{R}^D$; $\mathbf{D} \in \mathbb{R}^{D \times L}$; K or σ
2: Output: $\mathbf{c} \in \mathbb{R}^L$

3: $\mathcal{A} \leftarrow \{\}$; $\mathbf{c} \leftarrow \mathbf{0}$; $\mathbf{r} \leftarrow \mathbf{x}$
4: **while** $\|\mathbf{c}\|_0 \leq K$ **and** $\|\mathbf{r}\|_2 > \sigma$ **do**
5: $\quad \boldsymbol{\mu} \leftarrow \mathbf{D}^\top \mathbf{r}$
6: $\quad j^* \leftarrow \arg\max_j |\mu_j|, j \in \mathcal{A}^c$
7: $\quad \mathcal{A} \leftarrow \mathcal{A} \cup \{j^*\}$
8: $\quad \mathbf{c}_{\mathcal{A}} \leftarrow \left(\mathbf{D}_{\mathcal{A}}^\top \mathbf{D}_{\mathcal{A}}\right)^{-1} \mathbf{D}_{\mathcal{A}}^\top \mathbf{x}$
9: $\quad \mathbf{r} \leftarrow \mathbf{x} - \mathbf{D}\mathbf{c}$
10: **end while**

Each iteration t of the while-loop consists of two steps: atom selection and update of the coding vector. The atom that is most coherent to the current residual $\mathbf{r}^{(t-1)}$ is selected, and its index j^* is added to the active set of atoms $\mathcal{A}$ in lines 5 to 7. Then $\mathbf{c}^{(t)}$ is set to the coordinates of the orthogonal projection of $\mathbf{x}$ onto the subspace spanned by $\mathbf{D}_{(:,\mathcal{A})}$ in line 8, and the new residual $\mathbf{r}^{(t)}$ is computed in line 9. This procedure ensures that $\mathbf{r}^{(t)}$ is always orthogonal to the span of $\mathbf{D}_{(:,\mathcal{A})}$, and that the set of selected atoms is linearly independent. The algorithm terminates after $\|\mathbf{c}^{(t)}\|_0 = K$ or $\|\mathbf{x} - \mathbf{D}\mathbf{c}^{(t)}\|_2 \leq \sigma$ is reached. OMP converges with an exponential rate, which depends on the coherence of the dictionary to the signal class [24].

B. Basis Pursuit and LASSO

Convex relaxation of the ℓ_0 pseudo-norm to the ℓ_1 norm is another way to solve the sparse coding problem approximately. In this case, the cardinality constrained formulation (3) is known as the LASSO [38], and the error norm constrained formulation (4) is known as basis pursuit denoising [4]. Because both the objective function and the constraint are convex, the global minimum is unique and can be found efficiently using quadratic programming techniques.

In contrast to the ℓ_0 pseudo-norm, the ℓ_1 norm penalizes both the cardinality of the coding *and* the magnitude of the coding coefficients. The second property will be important for sparse coding in the concatenated dictionary, where large magnitude coefficients can lead to instabilities (see Section IV-B). If a sparse enough coding exists, the relaxed form is equivalent to the hard combinatorial problem, meaning that the sparsest coding can be found efficiently using the ℓ_1 norm [8].

1) Least Angle Regression (LARS): Least angle regression [11] is a very efficient iterative algorithm that obtains a solution closely resembling LASSO, and with a simple modification can be made to exactly obtain the LASSO solution. As with OMP, each iteration consists of an atom selection and a coding coefficient update step. Atom selection is identical to OMP. For the coefficient update, instead of an orthogonal projection onto the span of the selected atoms, LARS proceeds in the *equiangular* direction of the selected atoms, until a new atom has equal correlation with the residual as all atoms in the active set. Typically, either a cardinality-based or a norm-based stopping criterion is used (as with OMP).

2) Batch LARS With Coherence Criterion (LARC): We have extended LARS in two ways for our application. Both for dictionary learning and enhancement, a large number of observations have to be sparsely coded in the same dictionary. Precomputing the Gram matrix $\mathbf{G} = \mathbf{D}^\top \mathbf{D}$ avoids repeated computations of matrix-vector products involving $\mathbf{D}$, which are the most expensive operations in LARS. In addition, the matrix inverse $\mathbf{G}_{(\mathcal{A},\mathcal{A})}^{-1}$ is built iteratively using an update scheme based on the Cholesky factorization [32].

The second extension concerns the stopping criterion. As already mentioned in the introduction, an observation is a linear superposition of components, one of them coherent and one or more of them incoherent to the dictionary. Since only the coherent component can be sparsely coded in the dictionary, it can be separated from the other components by choosing the right value for the stopping criterion. Both OMP and LARS are greedy algorithms; therefore, the coherent components will be coded before the incoherent components, and the maximum residual coherence (line 6 in algorithm 1) will decrease with

every iteration [25]. This decrease suggests to use a residual coherence threshold μ_{dl} as the stopping criterion, which in contrast to the ℓ_2 norm of the residual or the ℓ_1 norm of the coefficient vector does not depend on the magnitude of the observation.

LARC (see algorithm 2) consists of three parts: atom selection (lines 6 to 8, computation of the equiangular direction $\mathbf{u}$ (lines 12 to 16) and update of $\mathbf{c}$ using the step size γ (lines 17 to 21). "$\min^+$" denotes that the minimum is only over positive arguments.

Note that the computation of $\mathbf{D}^\top\mathbf{r}^{(t-1)}$ in iteration t (line 6) is split into a constant part $\boldsymbol{\mu}^{(x)} = \mathbf{D}^\top\mathbf{x}$ and a variable part $\boldsymbol{\mu}^{(y)} = \mathbf{D}^\top\mathbf{y}^{(t-1)}$, which is updated efficiently in line 21. We have omitted the sequential Cholesky update of the matrix inverse $\mathbf{G}^{-1}_{(\mathcal{A},\mathcal{A})}$, the details are given, e.g., in [32].

Algorithm 2 Batch LARC

1: Input: $\mathbf{x} \in \mathbb{R}^D$; $\mathbf{D} \in \mathbb{R}^{D\times L}$; $\mathbf{G} = \mathbf{D}^\top\mathbf{D}$; μ_{dl}
2: Output: $\mathbf{c} \in \mathbb{R}^L$

3: $\mathbf{c} \leftarrow \mathbf{0}$; $\mathbf{y} \leftarrow \mathbf{0}$; $\mathcal{A} \leftarrow \{\}$
4: $\boldsymbol{\mu}^{(x)} \leftarrow \mathbf{D}^\top\mathbf{x}$; $\boldsymbol{\mu}^{(y)} \leftarrow \mathbf{0}$
5: **while** $|\mathcal{A}| < D$ **do**
6: $\quad\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}^{(x)} - \boldsymbol{\mu}^{(y)}$
7: $\quad j^* \leftarrow \arg\max_j |\mu_j|, j \in \mathcal{A}^c$
8: $\quad\mathcal{A} \leftarrow \mathcal{A} \cup \{j^*\}$
9: $\quad$**if** $|\mu_{j^*}| / \|\mathbf{x} - \mathbf{y}\|_2 < \mu_{\mathrm{dl}}$ **then**
10: $\quad\quad$**break**
11: $\quad$**end if**

12: $\quad\mathbf{s} \leftarrow \mathrm{sign}(\boldsymbol{\mu}_{\mathcal{A}})$
13: $\quad\mathbf{g} \leftarrow \mathbf{G}^{-1}_{(\mathcal{A},\mathcal{A})}\mathbf{s}$
14: $\quad b \leftarrow (\mathbf{g}^\top\mathbf{s})^{-1/2}$
15: $\quad\mathbf{w} \leftarrow b\mathbf{g}$
16: $\quad\mathbf{u} \leftarrow \mathbf{D}_{(:,\mathcal{A})}\mathbf{w}$

17: $\quad\mathbf{a} \leftarrow \mathbf{G}_{(:,\mathcal{A})}\mathbf{w}$
18: $\quad\gamma \leftarrow \min^+_{k\in\mathcal{A}^c} [(|\mu_{j^*}| - \mu_k)/(b - a_k), (|\mu_{j^*}| + \mu_k)/(b + a_k)]$
19: $\quad\mathbf{y} \leftarrow \mathbf{y} + \gamma\mathbf{u}$
20: $\quad\mathbf{c}_{\mathcal{A}} \leftarrow \mathbf{c}_{\mathcal{A}} + \gamma\mathbf{w}$
21: $\quad\boldsymbol{\mu}^{(y)} \leftarrow \boldsymbol{\mu}^{(y)} + \gamma\mathbf{a}$
22: **end while**

III. Dictionary Learning

An iterative dictionary learning algorithm adapts an initial dictionary to a particular signal class, such that observations from that signal class are sparsely coded in the dictionary with small error. As will be discussed in Section IV, such adaptation is typically necessary for successful enhancement, because the ideal dictionary has to be both coherent to its signal class as well as incoherent to all other signals present in the mixture.

Formally, a dictionary learning algorithm approximately factorizes a data matrix $\mathbf{X} \in \mathbb{R}^{D\times N}$ into a dictionary $\mathbf{D} \in \mathbb{R}^{D\times L}$ and a coding matrix $\mathbf{C} \in \mathbb{R}^{L\times N}$. The factorization optimizes the objective

$$\arg\min_{\mathbf{D},\mathbf{C}} \|\mathbf{X} - \mathbf{D}\cdot\mathbf{C}\|_F^2 \tag{5}$$

subject to a sparsity constraint on $\mathbf{C}$ and the unit ℓ_2 norm constraint on the atoms of $\mathbf{D}$. The approximation error is measured by the squared Frobenius norm $\|\cdot\|_F^2$, i.e., the sum of squares of all matrix elements. Due to the fact that both $\mathbf{D}$ and $\mathbf{C}$ are unknown, the objective function (5) is not convex, and the sparsity constraint on $\mathbf{C}$ makes finding the global optimum intractable. However, several authors (see [31] for a review) have proposed efficient algorithms based on an alternating minimization of $\mathbf{C}$ and $\mathbf{D}$ until convergence to a local optimum. We present the batch version of iterative dictionary learning used to train the speech and interferer dictionaries for our experiments. A productive implementation of the speech enhancement pipeline would additionally use an online algorithm [23] to learn and update the interferer dictionary.

A. Initialization

The initial dictionary $\mathbf{D}^{(0)}$ can be defined in various ways. For instance, the atoms can be chosen uniformly at random on the unit hypersphere, or can be sampled from the training data $\mathbf{X}$, followed by rescaling to unit length.

B. Coding Update

The squared Frobenius norm and the sparsity constraint are column separable, therefore minimizing the objective (5) for $\mathbf{C}$ given $\mathbf{D}$ amounts to N independent sparse coding problems, and in principle, any sparse coding algorithm (see Section II) is applicable. We use LARC in our experiments, i.e., given the previous dictionary $\mathbf{D}^{(t-1)}$, at iteration t of the dictionary learning algorithm each column $\mathbf{c}^{(t)}_{(:,n)},\ n = 1,\ldots,N$ of the coding matrix is updated as

$$\mathbf{c}^{(t)}_{(:,n)} \leftarrow \mathrm{LARC}\left(\mathbf{D}^{(t-1)},\, \mathbf{x}_{(:,n)},\, \mu_{\mathrm{dl}}\right).$$

The LARC sparse coding algorithm has three positive properties relevant to dictionary learning in our application. We have empirically observed that an ℓ_1 norm-based sparsity measure leads to better generalization performance of the dictionary than an ℓ_0 norm-based sparsity measure as employed by OMP (see below). Furthermore, the batch formulation of LARC is suited to solving a large number of independent coding problems given the same dictionary (see Section II-B2). Finally, setting the same residual coherence threshold for speech and all interferer types proved viable. A dictionary learning algorithm where the sparsity parameter had to be tuned would be unsuitable in our application, since the interferer properties are not known in advance.

C. Dictionary Update

The dictionary update step

$$\mathbf{D}^{(t)} \leftarrow \arg\min_{\mathbf{D}} \left\|\mathbf{X} - \mathbf{D}\cdot\mathbf{C}^{(t)}\right\|_F^2 \quad \text{s.t.}\ \left\|\mathbf{d}_{(:,l)}\right\|_2 = 1 \quad \forall l = 1,\ldots,L \tag{6}$$

is a least squares minimization problem with quadratic equality constraints, and can be solved e.g., using a Lagrangian dual formalism [19]. However, there is no need to find the exact minimum of (6), as long as the dictionary learning objective (5) is reduced in each iteration.

The K-SVD dictionary update step [1] is a greedy atom-by-atom update. For each atom $\mathbf{d}_{(:,l)},\ l = 1, \ldots, L$, the contribution to the residual norm due to $\mathbf{d}_{(:,l)}$ is isolated as

$$\begin{aligned} \|\mathbf{X} - \mathbf{D}\cdot\mathbf{C}\|_F^2 &= \left\| \mathbf{X} - \sum_{m=1}^{L} \mathbf{d}_{(:,m)}\mathbf{c}_{(m,:)} \right\|_F^2 \\ &= \left\| \left(\mathbf{X} - \sum_{m\neq l} \mathbf{d}_{(:,m)}\mathbf{c}_{(m,:)} \right) - \mathbf{d}_{(:,l)}\mathbf{c}_{(l,:)} \right\|_F^2 \\ &= \left\| \mathbf{R}^{(l)} - \mathbf{d}_{(:,l)}\mathbf{c}_{(l,:)} \right\|_F^2 . \end{aligned} \tag{7}$$

The residual norm is minimized using the one rank approximation $\mathbf{d}_{(:,l)}\mathbf{c}_{(l,\mathcal{N})}$ of $\mathbf{R}^{(l)}_{(:,\mathcal{N})}$, containing the columns $\mathcal{N} = \{n | C_{l,n} \neq 0, 1 \leq n \leq N\}$ of $\mathbf{R}^{(l)}$ where atom $\mathbf{d}_{(:,l)}$ was involved in the coding. The joint update of $\mathbf{d}_{(:,l)}$ and $\mathbf{c}_{(l,\mathcal{N})}$ ensures that the coding coefficients are adapted with respect to (w.r.t.) to the new atom, where $\mathcal{N}$ preserves the location of the nonzero coefficients in the coding row $\mathbf{c}_{(l,:)}$. The updated $\mathbf{d}_{(:,l)}$ and $\mathbf{c}_{(l,\mathcal{N})}$ are immediately available in the computation of $\mathbf{d}_{(:,l+1)}$ and $\mathbf{c}_{(l+1,\mathcal{N})}$.

To further reduce complexity, Rubinstein *et al.* [32] proposed a dictionary update step (algorithm 3) that approximates the SVD of $\mathbf{R}^{(l)}_{(:,\mathcal{N})}$ using power iterations. In our application, executing a single power iteration (lines 8 to 10) reduced the residual norm sufficiently.

Algorithm 3 Approximate K-SVD Dictionary Update

1: Input: $\mathbf{X} = \mathbb{R}^{D\times N}$; $\mathbf{D} = \mathbb{R}^{D\times L}$; $\mathbf{C} = \mathbb{R}^{L\times N}$
2: Output: Updated dictionary $\mathbf{D}$
3: **for** $l \leftarrow 1$ to L **do**
4: $\quad \mathbf{d}_{(:,l)} \leftarrow \mathbf{0}$
5: $\quad \mathcal{N} \leftarrow \{n | C_{l,n} \neq 0, 1 \leq n \leq N\}$
6: $\quad \mathbf{R} \leftarrow \mathbf{X}_{(:,\mathcal{N})} - \mathbf{D}\mathbf{C}_{(:,\mathcal{N})}$
7: $\quad \mathbf{g} \leftarrow \mathbf{c}^{\top}_{(l,\mathcal{N})}$
8: $\quad \mathbf{h} \leftarrow \mathbf{R}\mathbf{g}$
9: $\quad \mathbf{h} \leftarrow \mathbf{h}/\|\mathbf{h}\|_2$
10: $\quad \mathbf{g} \leftarrow \mathbf{R}^{\top}\mathbf{h}$
11: $\quad \mathbf{d}_{(:,l)} \leftarrow \mathbf{h}$
12: $\quad \mathbf{c}_{(l,\mathcal{N})} \leftarrow \mathbf{g}^{\top}$
13: **end for**

The success of dictionary learning is measured by the ability of a trained dictionary to achieve a low approximation error when sparsely coding test data not seen during training. Coding a test observation with increasing cardinality K, the use of a coherent dictionary results in a rapid decay of the approximation error, compared to the slow decay of using an incoherent dictionary that only yields a low approximation error when K approaches D.

IV. Speech Enhancement

We model each observed frame $\mathbf{x}^{(n)} \in \mathbb{R}^D$ of degraded speech as a linear additive mixture

$$\mathbf{x}^{(n)} = \mathbf{s}^{(n)} + \mathbf{i}^{(n)} \tag{8}$$

of target speech frame $\mathbf{s}^{(n)} \in \mathbb{R}^D$ and interferer frame $\mathbf{i}^{(n)} \in \mathbb{R}^D$. Given a single frame $\mathbf{x}$ (where the dependency on n is from now on omitted for notational clarity), speech enhancement pursues the goal to obtain an estimate $\hat{\mathbf{s}}$ of the underlying clean speech signal such that the residual norm $\|\hat{\mathbf{s}} - \mathbf{s}\|_2$ is significantly lower than $\|\mathbf{x} - \mathbf{s}\|_2$ (see Section VI-B for further discussion of measuring enhancement performance).

Speech enhancement is successful if the speech dictionary $\mathbf{D}^{(s)}$ is coherent to the speech signal *and* incoherent to the interferer signal. Classical coherent denoising considers the case where the interferer is pure noise (e.g., zero-mean Gaussian white noise) and does not contain any structure. Such an interferer is incoherent to any fixed dictionary [28], and in particular to the speech dictionary. This case is covered in Section IV-A.

Many relevant kinds of interferers contain structure. If the structured component of the interferer signal is also incoherent to the speech dictionary, the treatment is equivalent to the unstructured case. If the interferer is partially coherent to the speech dictionary, there is a risk that parts of the interferer signal will be confused as coming from the speech source, but for a structured interferer, training a coherent interferer dictionary $\mathbf{D}^{(i)}$ is possible. A sparse coding of the degraded speech observation in the *composite dictionary* $\mathbf{D} = \left[\mathbf{D}^{(s)}\, \mathbf{D}^{(i)}\right]$ significantly improves enhancement performance, if $\mathbf{D}^{(s)}$ is more coherent to $\mathbf{s}$ than $\mathbf{D}^{(i)}$, and $\mathbf{D}^{(i)}$ is more coherent to $\mathbf{i}$ than to $\mathbf{D}^{(s)}$. This more general case is covered in Section IV-B.

A. Incoherent Interferers

As mentioned, unstructured interferers cannot be sparsely represented in any fixed dictionary, in particular also not in a speech dictionary. As a prominent example for an incoherent interferer scenario, we consider the enhancement of speech degraded by zero-mean Gaussian white noise. In the enhancement step, the degraded speech mixture is sparsely coded in the speech dictionary using LARC with a suitably chosen residual coherence threshold μ_{enh}. LARC coding captures the structured speech signal components which have a coherence to the speech dictionary that is above the threshold, while discarding the interferer components, as they fall below the residual coherence threshold.

Formally, an observation $\mathbf{x}$ of degraded speech is sparsely coded in the speech dictionary $\mathbf{D}^{(s)}$ using LARC with a residual coherence threshold μ_{enh}, to obtain the vector of coding coefficients $\mathbf{c}^{(s)} \in \mathbb{R}^{L_s}$:

$$\mathbf{c}^{(s)} \leftarrow \text{LARC}\left(\mathbf{D}^{(s)}, \mathbf{x}, \mu_{\text{enh}}\right).$$

Coding with a suitable μ_{enh} leads to a coding vector $\mathbf{c}^{(s)}$ where large weights explain speech contributions in the mixture $\mathbf{x}$

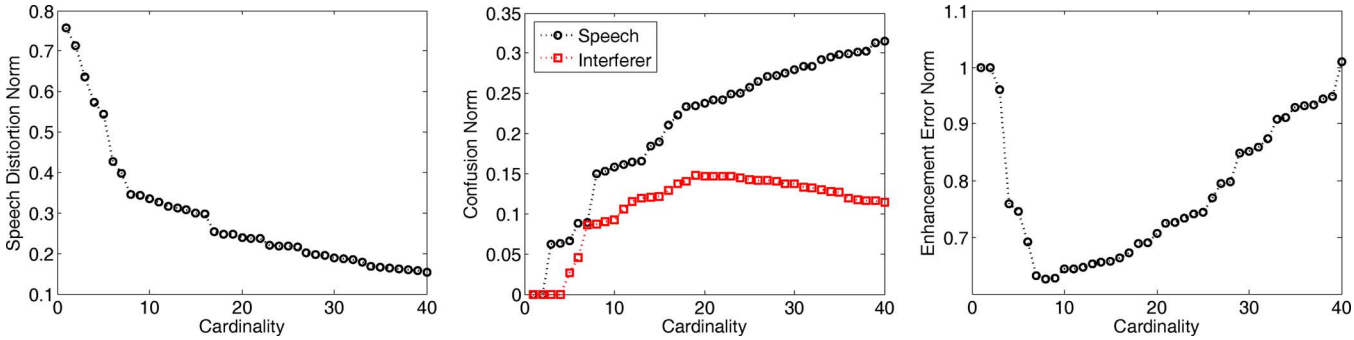


Fig. 2. Illustration of the tradeoff between source distortion and source confusion, for LARC coding a single frame of clean speech degraded by babble noise. The *left* figure plots the speech distortion error $\|\mathbf{s}-\hat{\mathbf{s}}\|_2$ for coding $\mathbf{s}$ in $\mathbf{D}^{(s)}$, which steadily decreases for increasing coding cardinality $\|\mathbf{c}^{(s)}\|_0$. The *middle* figure plots the source confusion error, measured by $\|\hat{\mathbf{s}}\|_2$ when coding $\mathbf{i}$ in the composite dictionary $\mathbf{D}$ ("Speech" curve), and $\|\hat{\mathbf{i}}\|_2$ when coding $\mathbf{s}$ in $\mathbf{D}$ ("Interferer" curve). Note that for the latter, no confusion occurs for $\|\mathbf{c}\|_0 < 5$. The *right* figure plots the enhancement error $\|\mathbf{s}-\hat{\mathbf{s}}\|_2$ for coding $\mathbf{x}$ in $\mathbf{D}$, which is minimal for $\|\mathbf{c}\|_0 = 8$. For a sparser coding, the error is dominated by source distortion, whereas for a denser coding, the error is dominated by source confusion.

using atoms from the speech dictionary $\mathbf{D}^{(s)}$. An estimate of the underlying clean speech is obtained as $\hat{\mathbf{s}} = \mathbf{D}^{(s)}\mathbf{c}^{(s)}$.

The above method provides excellent results as long as the considered interferer is sufficiently incoherent to the speech dictionary [24, ch. 12]. However, the more coherent an interferer becomes to the speech dictionary, the more likely it is that interferer components in the residual are explained by speech dictionary atoms, instead of being discarded by falling below μ_{enh}. For interferers which are partially coherent to the speech dictionary and also contain structure, a better approach is possible, based on sparse coding in the composite dictionary consisting of a speech and an interferer dictionary. This is discussed in the next section.

B. Partially Coherent Interferers

A structured interferer can be sparsely represented with low approximation error in a suitably trained dictionary. In order to enhance speech degraded by structured interferers which are partially coherent to the speech dictionary, the degraded speech mixture is sparsely coded in the composite dictionary consisting of the concatenation of the speech and the interferer dictionary.

Formally, an observation $\mathbf{x}$ of degraded speech is sparsely coded in the composite dictionary $\mathbf{D} = [\mathbf{D}^{(s)}\,\mathbf{D}^{(i)}]$ using LARC with a residual coherence threshold μ_{enh}, to obtain the coding vector $\mathbf{c}$:

$$\mathbf{c} \leftarrow \text{LARC}\left([\mathbf{D}^{(s)}\,\mathbf{D}^{(i)}],\, \mathbf{x},\, \mu_{\text{enh}}\right).$$

The vector $\mathbf{c} = [\mathbf{c}^{(s)};\, \mathbf{c}^{(i)}]$ (concatenation in column-direction) consists of weights $\mathbf{c}^{(s)}$ corresponding to the speech dictionary $\mathbf{D}^{(s)}$, as well as weights $\mathbf{c}^{(i)} \in \mathbb{R}^{L_i}$ corresponding to the interferer dictionary $\mathbf{D}^{(i)}$. An estimate of the underlying clean speech is again obtained as $\hat{\mathbf{s}} = \mathbf{D}^{(s)}\mathbf{c}^{(s)}$.

An *exact recovery condition (ERC)* [10] establishes the requirements on the signals and dictionaries, such that the speech contribution to the mixture is explained by weights in $\mathbf{c}^{(s)}$ only, and the structured interferer contribution is explained by weights in $\mathbf{c}^{(i)}$ only, i.e., no confusions occur between the sources. Assuming that both the speech and interferer signals have been generated by their respective dictionaries, i.e., $\mathbf{s} = \mathbf{D}^{(s)}\tilde{\mathbf{c}}^{(s)}$ and $\mathbf{i} = \mathbf{D}^{(i)}\tilde{\mathbf{c}}^{(i)}$, Donoho and Huo [10] show that recovery is exact if

$$\|\tilde{\mathbf{c}}^{(s)}\|_0 + \|\tilde{\mathbf{c}}^{(i)}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}^{(s)}, \mathbf{D}^{(i)})}\right) \tag{9}$$

where the *mutual coherence* between the speech and interferer dictionary is defined as

$$\mu(\mathbf{D}^{(s)}, \mathbf{D}^{(i)}) = \max_{p,q} \left|\mathbf{d}^{(s)\top}_{(:,p)}\mathbf{d}^{(i)}_{(:,q)}\right| \tag{10}$$

with $p = 1, \ldots, L_s$ and $q = 1, \ldots L_i$ (also see [8] for the case of an additional unstructured component in the mixture). Therefore, if the coding is sparse enough or the dictionaries are incoherent enough, no source confusion occurs.

For partially coherent interferers like speech babble, the requirements of the ERC are not satisfied, but (9) still motivates how varying $\|\mathbf{c}\|_0$ (by varying μ_{enh}) controls the tradeoff between source distortion and source confusion errors that both contribute to the estimation error $\|\mathbf{s}-\hat{\mathbf{s}}\|_2$ (see Fig. 2 for an illustration). For a very sparse $\mathbf{c}$, source confusions are unlikely according to (9) as long as $\mu(\mathbf{D}^{(s)}, \mathbf{D}^{(i)})$ is small enough, but $\|\mathbf{x} - \mathbf{D}\mathbf{c}\|_2$ is expected to be large because $\mathbf{c}$ lacks the necessary degrees of freedom in the approximation of $\mathbf{x}$. Consequentially, the clean speech estimate will sound distorted, due to a too sparse coding of the speech source. On the other hand, $\|\mathbf{x} - \mathbf{D}\mathbf{c}\|_2$ can be made arbitrarily small by increasing $\|\mathbf{c}\|_0$. However, (9) predicts that source confusions become more likely, which become apparent when separating $\mathbf{D}\mathbf{c}$ into $\mathbf{D}^{(s)}\mathbf{c}^{(s)}$ and $\mathbf{D}^{(i)}\mathbf{c}^{(i)}$. The resulting clean speech estimate will have insufficient interferer attenuation.

In our experiments, we have consistently observed that both source distortion and source confusion errors grow gradually with an increasing violation of the ERC, and that a tradeoff is possible which leads to significantly better enhancement performance than an extremely sparse or dense coding. Furthermore, using LARC instead of OMP in the enhancement step improved the temporal smoothness of the enhanced speech, since the ℓ_0 pseudo-norm (as used in OMP) does not penalize the magnitudes of the coding coefficients. If the optimal coding cardinality is not very sparse, the OMP algorithm can show numerical instability due to insufficient regularization. It then explains the residual using a combination of nearly colinear atoms

with coding coefficients that have large magnitudes and opposite signs. After separating the code $\mathbf{c}$ into the speech and interferer contributions $\mathbf{c}^{(s)}$ and $\mathbf{c}^{(i)}$, the large magnitude coding coefficients no longer cancel each other, if the corresponding atoms were chosen from both the speech and interferer dictionaries. This issue becomes audible in the separated signals as temporal fluctuations. The ℓ_1 norm of LARC penalizes both the cardinality and the coefficient magnitude of $\mathbf{c}$, and therefore, the method achieves better stability due to a more effective regularization.

V. Related Work

Our speech enhancement method employs both target speech and interferer models. We evaluate the performance of our approach against two other model-based enhancement approaches, that differ in model complexity and assumptions about the nature of the speech and interferer signals.

A. Geometric Spectral Subtraction

Geometric spectral subtraction (GA) was proposed in [21] to address the problem of residual *musical noise* encountered in standard spectral subtraction. As discussed in Section I-A, a spectral subtraction algorithm estimates the average interferer spectrum during speech inactivity, and subtracts it from the mixture spectrum during speech activity. A naive subtraction cannot perfectly recover the speech spectrum in nonstationary interferer scenarios; however, as at times it subtracts too little or too much from the mixture spectrum, producing random isolated peaks of residual interferer energy.

This fact follows from transforming (1) into the power spectral domain:

$$|X(\omega,n)|^2 = |S(\omega,n)|^2 + |I(\omega,n)|^2 + S(\omega,n)I^*(\omega,n) + S^*(\omega,n)I(\omega,n) \quad (11)$$

where $X(\omega,n)$, $S(\omega,n)$ and $I(\omega,n)$ denote the complex mixture, speech and interferer short-time Fourier spectra, respectively, and $I^*(\omega,n)$ denotes the complex conjugate of $I(\omega,n)$. Since the speech and the interferer are assumed to be statistically independent, the cross terms in (11) vanish in expectation. However, for short intervals they do not vanish, and omitting them introduces an approximation error. As shown in [21], this error is most severe if the SIR in a frequency bin is close to 0 dB, which is often the case in speech enhancement applications.

Lu and Loizou therefore proposed short-term estimators for both the speech and the interferer spectrum. Using geometric arguments, a suppression rule

$$\left|\hat{S}(\omega,n)\right| = H_{\mathrm{GA}}(\omega,n)\,|X(\omega,n)|$$

for estimating the clean speech magnitude spectrum at frame n was derived, with

$$H_{\mathrm{GA}}(\omega,n) = \frac{|S(\omega,n)|}{|X(\omega,n)|} = \sqrt{\frac{1-\frac{[\gamma(\omega,n)+1-\xi(\omega,n)]^2}{4\gamma(\omega,n)}}{1-\frac{[\gamma(\omega,n)-1-\xi(\omega,n)]^2}{4\xi(\omega,n)}}} \quad (12)$$

where $\xi(\omega,n)$ is the *instantaneous a priori SNR* and $\gamma(\omega,n)$ is the *instantaneous a posteriori SNR*

$$\xi(\omega,n) = \frac{|S(\omega,n)|^2}{|I(\omega,n)|^2}, \quad \gamma(\omega,n) = \frac{|X(\omega,n)|^2}{|I(\omega,n)|^2}. \quad (13)$$

In contrast to naive spectral subtraction, the above rule (12) does not assume that the cross terms in (11) vanish. At each frame n, $\xi(\omega,n)$, and $\gamma(\omega,n)$ are recursively computed from the clean speech spectrum estimate of past frames and the interferer spectrum estimate, which is initialized with the average interferer spectrum obtained from a speech pause, and continuously updated using minimum statistics [26]. The suppression weights $H_{\mathrm{GA}}(\omega,n)$ are limited to 1. Further details including smoothing and thresholding constants are given in [21].

Our approach is conceptually similar to geometric spectral subtraction in that both algorithms enhance the mixture based on instantaneous estimates of the speech and interferer spectra. However, GA does not incorporate an explicit speech model, and therefore its performance crucially depends on how well the interferer spectrum is tracked by the minimum statistics and recursive estimation approach.

B. Codebook-Based Spectral Filtering

Several authors have trained speech and interferer models using vector quantization. For example, Ellis and Weiss [13] trained a speaker dependent codebook in the complex STFT domain. The codebook vectors take into account both magnitude and phase of the Fourier coefficients, but the phase change was quantized instead of the absolute phase. Lacking an interferer model, the enhancement step consisted of projecting the mixture frame on the closest clean speech prototype of the trained codebook. Furthermore, to exploit temporal constraints of the speech source, a discrete hidden Markov model (HMM) was trained on top of the clean speech codebook. The authors tested their approach by separating speech from speech shaped noise, but for the example given in the paper (0-dB mixture SIR) the method failed to improve the mixture signal. They noted three problems with their enhancement approach: the codebook had to be very large to reach an acceptably low source distortion, which is to be expected given a one-sparse coding of speech. Second, considerable interferer energy remained in speech pauses after the projection step. The constant one-sparse coding forced an explanation of pure interferer frames with elements of the speech codebook, which might fit quite well if the interferer has speech like characteristics. Finally, including the HMM did not improve performance, suggesting that the constant one-sparse coding of the mixture in the speech codebook fundamentally limits the performance of this approach.

Srinivasan *et al.* [36] trained both speech and interferer codebooks by means of vector quantization in the LPC domain, using Itakura–Saito (IS) distortion as the distance measure. The long-term interferer spectrum (estimated by minimum statistics) was also included in the interferer codebook. In the enhancement step, the best matching pair of speech and interferer spectral shapes (with associated gains) was estimated from the mixture by minimizing the IS distortion between the mixture power spectrum and the linear combination of the

speech and the interferer spectrum. Instead of a full optimization over all possible combinations of speech and interferer codebook elements, Srinivasan *et al.* proposed a greedy search. Finally, a Wiener filter was constructed from the spectral shapes and associated gains.

The interferer codebook size had to be chosen with care. A too generic interferer codebook turned out to be problematic, because the optimal pair of speech and interferer spectra became increasingly ambiguous. Therefore, the authors proposed to train several interferer specific codebooks, and to select the one appropriate for enhancement based on a classification of the long-term interferer spectrum. Furthermore, the LP order and the codebook size was optimized for each interferer type. While avoiding the complexity of using a VAD to obtain interferer training data, this approach presupposes that an appropriate interferer codebook is available for enhancement.

Srinivasan *et al.* proposed two different approaches to reduce the distortion resulting from the one-sparse coding of speech. An interpolation codebook was proposed in [36], and in [37] the mixture was explained by a Bayesian averaging over all pairs of speech and interferer codebook vectors. The full Bayesian treatment was simplified by assuming that the spectral shapes and gains are all independent, and that the likelihood is strongly peaked around the maximum-likelihood (ML) values for the speech and interferer gains. This assumption resulted in a weighted average over all pairs of vectors from the speech and interferer codebooks, where the weight of each pair is the product of the prior probabilities of the spectra.

Dictionary learning generalizes vector quantization [1], in that both speech and interferer are approximated by a sparse linear combination of multiple atoms, instead of a single codebook vector. This relaxation leads to a significant reduction of the source distortion error. Instead of also re-implementing the different feature space, the distance measure and the interpolation scheme of [36], our evaluation only focuses on the improvement gained by sparse coding with cardinality greater than one in both the dictionary learning and the enhancement steps. We therefore evaluate vector quantization and dictionary learning based enhancement in the same feature space, and compare our method to codebook-based spectral filtering where a mixture observation is explained using a weighted linear combination of one vector from the speech codebook and one vector from the interferer codebook.

VI. Evaluation

We evaluate the performance of our method in comparison with established and powerful baselines, both in the speaker dependent and speaker independent case, with interferer signals obtained from a range of relevant environments. As the performance of speech enhancement strongly depends on the tradeoff between source distortion and source confusion, we give insights on how to balance the two effects for optimal enhancement results. We also discuss design choices regarding feature extraction, dictionary learning, enhancement, and resynthesis.

For the purpose of evaluating the performance of our enhancement approach, an environment specific interferer dictionary was trained for each interferer class. The ability of an interferer dictionary to generalize to signals from another interferer class is not evaluated, as environment specific interferer observations are assumed to be obtainable during speech pauses (as discussed in Section I).

A. Data

Speech data is obtained from the GRID[2] audio-visual corpus, which provides a total of 34 speakers of both genders. We have used the audio data of 15 male and 15 female speakers in our experiments. For each speaker, the corpus contains 1000 sentences of a simple grammatical structure without high-level linguistic cues, for instance "Place green at B 4 now" [5].

As nonstationary interferer data, we used location recordings made in the following environments: classical *piano* music replayed in a living room, *street* traffic noise and *wind* noise of an exposed microphone, all obtained from a proprietary corpus. Furthermore, obtained from the NOISEX-92 corpus [40]: speech *babble* noise, machine noise in a *factory* and engine and tire noise in a *Volvo* car. As a maximally unstructured and non-sparse interferer, synthetic zero-mean Gaussian *white* noise was used as well. Both the speech and the interferer data was sampled down to 16 kHz.

Dictionary learning as well as parameter optimization for dictionary learning and speech enhancement was performed on a training and a test data set. Final enhancement performance is reported on a separate validation set. In the speaker independent case, the speech dictionary was trained on six male and six female speakers of the training set, and the enhancement performance was averaged over the validation data of three male and three female speakers of the validation set. In the speaker-dependent case, a dictionary was inferred on training data for each speaker of the validation set, and the enhancement performance was established on separate validation data of that speaker. The final results are reported as the average performance over all validation speakers. This procedure guarantees that the enhancement performance is established on the same validation data in the speaker dependent and independent case.

For the speech data, splitting was performed on a file-by-file basis. For the interferer data, the recordings were each split into three disjoint segments for training, test and validation, and an interferer specific dictionary was trained for each interferer class on the respective training data.

B. Performance Measures

Speech enhancement algorithms aim to improve both the speech *quality* and the speech *intelligibility*. A high-quality speech signal is perceived as being natural and pleasant to listen to, and free of distracting artifacts. However, measuring speech quality is challenging, as it is subjective. Speech intelligibility on the other hand is measured by word error rates in a speech recognition scenario. Good speech quality does not necessarily imply good speech intelligibility, and vice-versa. For instance, low-quality synthesized speech can be highly intelligible.

The performance of a speech enhancement algorithm can be measured both *subjectively* and *objectively*. Subjective measurements are based on the judgment of human listeners, and are important because in many applications (such as hearing

[2]http://www.dcs.shef.ac.uk/spandh/gridcorpus/

aids) the output of the enhancement algorithm has to appeal to the human ear. However, subjective evaluation takes time, is expensive, and usually requires trained listeners. As an alternative, objective measures provide mathematical models of some perceptual aspects of the human auditory system.

We report results measured in the *frequency-weighted segmental SNR (fwSegSNR)* and in *PESQ scores*, which were shown to correlate well with both subjective speech quality and subjective speech intelligibility scores [22]. fwSegSNR is a conceptually simple objective measure, computed on individual signal frames, and the per-frame scores are averaged over time. We use the frequency-domain definition of the measure, incorporating a perceptually motivated frequency-band weighting as well as frequency-band spacing.

Given the true and estimated speech magnitude spectra, the frequency-weighted segmental SNR is defined as

$$\frac{10}{N\bar{w}}\sum_{n=1}^{N}\sum_{b=1}^{B} w_b \log_{10}\frac{|S(b,n)|^2}{\left(|S(b,n)|-|\hat{S}(b,n)|\right)^2} \tag{14}$$

where $S(b,n)$ is the frequency-domain representation of the clean speech signal, for frequency band b and time frame n, $\hat{S}(b,n)$ is the frequency-domain representation of the estimated speech signal, N is the total number of frames, B is the total number of frequency bands, w_b is the weight of frequency-band b, and $\bar{w}=\sum_{b=1}^{B} w_b$. The frequency-band weights w_b are based on the articulation index [20]. We use the Matlab implementation provided by [20].

The computation of PESQ scores is considerably more involved, the details are given in [29]. The performance results expressed in PESQ scores are available from the authors' website. We again use the Matlab implementation provided by [20].

C. Feature Space

Speech enhancement was performed in the STFT magnitude domain and the MDCT domain. The MDCT is a real valued transform that includes phase information, for the STFT we only consider the magnitude and omit the phase, which implies that (2) holds only approximately in this case. Other representations could be chosen, as long as the feature extraction is (approximately) linear and the distance function is perceptually meaningful and mathematically tractable. We have found the ℓ_2 distance to correlate quite well with perceptually motivated measures for our choice of feature spaces.

The number of frequency bins per frame is determined by the length of the time-domain analysis window, where a Hamming window was chosen for the STFT and a Kaiser–Bessel derived window was chosen for the MDCT. The temporal smoothness of frames is determined by the time-domain analysis window overlap, where a minimum amount of overlap is necessary to avoid aliasing. The transform domain was tiled in overlapping blocks (see Fig. 3), where the block height specifies the number of frequency bins, and the block width specifies the number of consecutive frames. A tall and narrow block (A in Fig. 3) captures more of the harmonic content, whereas a short and wide block (B in Fig. 3) captures more of the temporal dynamics. The final feature space for dictionary learning and enhancement is based on vectorized blocks.

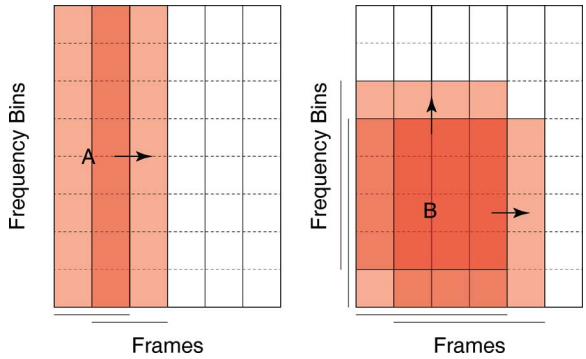


Fig. 3. The figure on the left shows a scenario where the transform domain is tiled using tall and narrow blocks (A). In this case, the blocks are overlapped and shifted only on the time axis, since the frequency axis is fully covered by each block. The figure on the right shows a scenario where tiling occurs using a short and wide block (B). In this case, the blocks are overlapped and shifted both on the frequency and the time axis. For the same number of bins, a tall and narrow block favors spectral information, whereas a short and wide block favors temporal dynamics of the signal.

As discussed in Section IV, estimating the clean speech from the mixture is possible if the speech dictionary is coherent to the target and incoherent to the interferer. Maximizing the coherence of the dictionary to its signal class is more easily achieved in a low-dimensional feature space, i.e., a short analysis window or short block height and a single analysis frame per block. For structured signals, this results in low source distortion even for very sparse codings. Conversely, maximizing the incoherence of the dictionary to all other signal classes requires a high-dimensional feature space. This minimizes source confusion and is achieved by increasing the block height, block width, or both.

We have investigated analysis window lengths of up to 64 ms, and used a constant analysis window shift of 10 ms. The block height was varied from 8 frequency bins to full height, and the block width was varied from two to 64 frames. The resulting feature space dimensionality varied from 512 up to 2048 dimensions.

D. Dictionary Learning

An iterative dictionary learning algorithm is specified by the sparse coding algorithm, the dictionary update method and the choice of the initial dictionary. We used LARC (algorithm 2) as the sparse coding method, and the computationally efficient approximate SVD atom update step (algorithm 3) of [32]. Since the interferer properties are not known in advance, a dictionary learning algorithm where the sparsity parameter had to be tuned would be unsuitable. Specifying the desired coding sparsity using the residual coherence threshold μ_{dl} has the advantage over a fixed coding cardinality K that incoherent signal frames (such as background noise in the speech corpus) are rejected, and has the advantage over setting a residual noise variance σ that adaptation to the energy distribution of the signal is not necessary. We have investigated residual coherence thresholds from 0.2 to 0.6 for the sparse coding step.

We evaluated two initialization methods: the atoms of the initial dictionary were either sampled uniformly on the unit hypersphere, or obtained by resampling the training data. The first

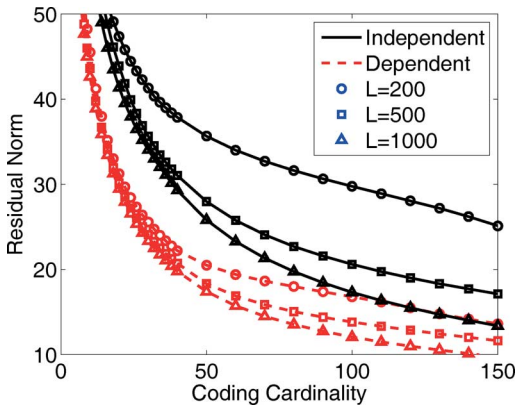


Fig. 4. Generalization performance of speaker-independent (solid black) and dependent (dashed red) dictionaries on validation data ($D = 560$), for dictionary sizes of L =200, 500, and 1000 atoms. Better performance can be achieved in the speaker dependent case, which is indicated by a faster decay of the median residual norm. Increasing the number of atoms improves coherence of the dictionary to its signal class in both cases.

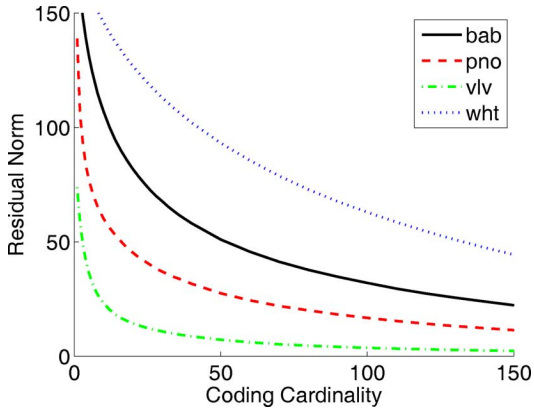


Fig. 5. Generalization performance of interferer dictionaries ($L = 2000$) on validation data ($D = 560$), for babble, piano, car, and white noise interferers (all having the same average power). The rate of decay of the median residual norm, in comparison to Gaussian white noise, indicates the amount of structure present in the signal which is amenable to dictionary learning.

method generates an initial dictionary that is not adapted to the training data at all, whereas the second method results in atoms which are overly adapted to single observations. We have consistently observed that the latter sampling scheme leads to significantly faster convergence and better generalization performance than random initialization of the dictionary.

The algorithm approximately converged after twenty iterations of coding and dictionary updates.

1) Speech Dictionary: Both for the speaker dependent and independent case, the training utterances were randomly sub-sampled on a file and block level to obtain five minutes of clean speech. We have trained dictionaries containing up to 2000 atoms. Fig. 4 illustrates the generalization performance of the speech dictionary, measured by the residual norm versus coding cardinality curve, in the speaker independent and dependent case. Increasing the dictionary size leads to lower residual norm for a given cardinality, with diminishing returns for larger dictionaries. Furthermore, better coherence of the dictionary to the signal class is achieved in the speaker dependent case, where the 200 atom dictionary outperforms the 1000 atom speaker independent dictionary. This is to be expected, given the greater spectral variation induced by the different genders in the speaker-independent case.

2) Interferer Dictionary: We have trained dictionaries on a consecutive 30-second segment for each interferer. The algorithm parameters were identical for all interferers, and set to the same values as in speech dictionary learning. Plotting the generalization performance (see Fig. 5) reveals the amount of structure present in each interferer class. Compared to unstructured Gaussian white noise, a rapid decay of the median residual norm is achieved for car noise and piano music, indicating that there is prominent structure which is effectively modeled by a coherent trained dictionary. The generalization performance of the interferer dictionary is also predictive for the overall speech enhancement performance (see Section VI-E): a larger improvement is achieved both for highly structured and unstructured interferers, whereas a smaller improvement is achieved for the speech babble interferer, which contains an intermediate amount of structure that is partially coherent to the speech signal.

E. Speech Enhancement

The enhancement performance of all methods are evaluated using synthetic mixtures, which are generated by linear additive mixing of clean speech and interferer signals in the time-domain at various speech to interferer power ratios (SIRs). For the computation of the SIR, the signals were preprocessed with an A-weighted filter, in order to model the auditory sensitivity to different frequencies. An improvement is deemed significant if the median fwSegSNR value given the enhanced and the clean speech is above the 75th percentile of the fwSegSNR value given the degraded and the clean speech.

For the STFT domain, the speech magnitude estimate can be combined with the mixture phase, or the speech and interferer magnitudes provide estimates of the instantaneous *a priori* and *a posteriori* SNR, from which a suppression rule can be derived for filtering the mixture magnitude. We used the suppression rule of (12) in our experiments both for codebook-based filtering and our proposed algorithm, which enables a direct comparison with geometric spectral subtraction. Binary masking or Wiener filtering would be other options.

For all interferers and at all mixture SIRs that were tested, enhancement in the STFT magnitude domain significantly outperformed enhancement in the MDCT domain. Although (8) only holds in approximation for the STFT magnitude domain, and resynthesis of the time domain signal relies on the degraded mixture phase, the additional complexity of modeling phase information in the MDCT proved to have a greater negative impact on enhancement performance. We, therefore, only report detailed results for the STFT magnitude features.

In Figs. 6 and 7, *X* denotes the fwSegSNR value given the degraded speech and the clean speech, i.e., the objective measurement before any enhancement. Note that the fwSegSNR measure is sensitive to the spectral characteristics of each interferer class. This sensitivity explains the fact that for a fixed mixture SIR (e.g., 10 dB), the frequency-weighted segmental SNR shows different values for different interferers. Furthermore, *GA* denotes geometric spectral subtraction, *VQ* denotes codebook-

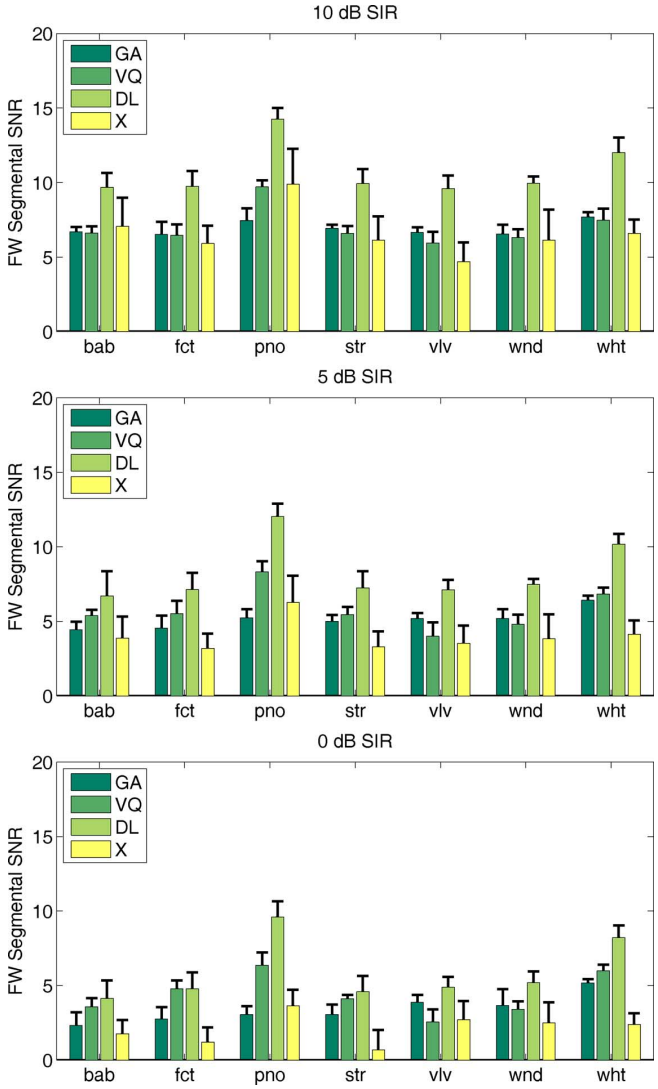


Fig. 6. Enhancement performance in the speaker *independent* case at three different *signal-to-interferer ratios* (SIR), for seven different interferers, speech babble (*bab*), factory noise (*fct*), piano music (*pno*), street noise (*str*), engine and tire noise (*vlv*), wind noise (*wnd*), and Gaussian white noise (*wht*). *GA* denotes geometric spectral subtraction, *VQ* denotes codebook-based filtering, *DL* denotes our method and *X* denotes the objective measurement before any enhancement. The median fwSegSNR value is denoted by the filled bar, while the whisker denotes the 75th percentile of the distribution of fwSegSNR values.

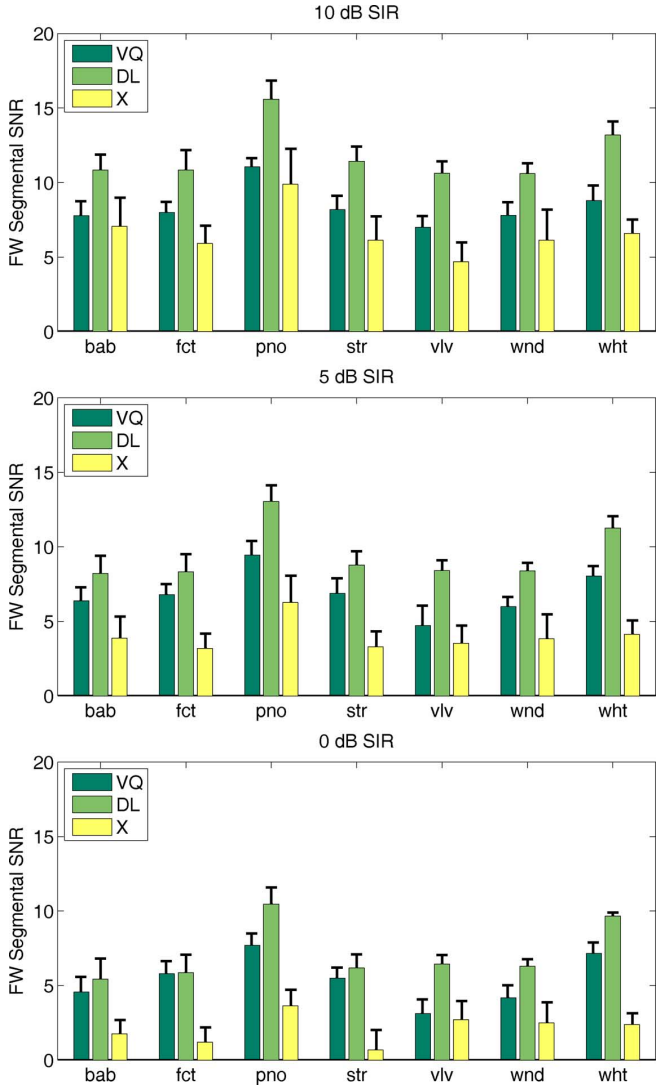


Fig. 7. Enhancement performance in the speaker *dependent* case at three different *signal-to-interferer ratios* (SIR), for seven different interferers, speech babble (*bab*), factory noise (*fct*), piano music (*pno*), street noise (*str*), engine and tire noise (*vlv*), wind noise (*wnd*), and Gaussian white noise (*wht*). *GA* denotes geometric spectral subtraction, *VQ* denotes codebook-based filtering, *DL* denotes our method and *X* denotes the objective measurement before any enhancement. The median fwSegSNR value is denoted by the filled bar, while the whisker denotes the 75th percentile of the distribution of fwSegSNR values.

based filtering using a linear combination of exactly one atom from the speech dictionary and one atom from the interferer dictionary (see Section V-B), and *DL* denotes our method.

The codebook size and the dictionary size were set to $L = 1000$ atoms each for the speech and the interferer dictionary. For dictionary learning, the residual coherence threshold μ_{dl} was set to 0.2 for all dictionaries. μ_{enh} during enhancement was set to 0.15 and 0.1 in the speaker dependent and independent case, respectively. Both in the speaker dependent and speaker independent case, the maximum analysis window length (64 ms), full block height and maximum feature space dimensionality (2048) was optimal for VQ and DL. Given the range of block heights and widths that were tested, spectral information proved to be more important than temporal dynamics, i.e., tall and narrow blocks led to better results than short and wide blocks (for the same feature space dimensionality). Further analysis did show that although coherent dictionaries can also be trained for wide blocks, tall blocks result in smaller confusion error, which implies that all else being equal, the speech contribution to the mixture can be better distinguished from the interferer contribution based on spectral than based on temporal information.

We have also evaluated a variant of our method where only the speech dictionary (instead of the composite dictionary) was used in the enhancement step. As expected from the discussion in Sections IV-A and IV-B, a significant improvement was only achieved for the white noise interferer; therefore, we omit detailed results.

In the speaker independent case (Fig. 6), DL significantly outperforms the comparison methods in all interferer scenarios at 10-dB and 5-dB SIR, and in 5 out of 7 interferer scenarios at 0-dB SIR. Our method achieves a median fwSegSNR gain of 3.6 dB at 10-dB SIR, a median gain of 3.8 dB at 5-dB SIR, and

a median gain of 3.2 dB at 0-dB SIR. In the case of lightly degraded speech (10-dB SIR), DL achieves a significant improvement for all interferer scenarios, whereas GA and VQ can introduce strong artifacts which further degrade the mixture signal. Note that our method shows the highest gain for the piano and the white noise interferer. This gain is explained by the fact that the piano interferer is very structured, which enables the learning of a good interferer model. On the other hand, the white noise interferer is unstructured and incoherent to speech, and is thus disregarded by LARC coding.

In the speaker dependent case (Fig. 7), both DL and VQ achieve higher performance gains than in the speaker independent case. This is due to the fact that speaker dependency allows for a more coherent speech dictionary (cf. Fig. 4). At 10-dB SIR, DL achieves a median performance gain of 4.8 dB, where at 5-dB SIR a median gain of 5.2 dB and at 0-dB SIR a median gain of 4.6 dB is achieved. The performance increase from VQ to DL is smaller in the speaker dependent case, because less source distortion is introduced by the maximally sparse coding when the codebook is more coherent to the speech source. However, the increase from VQ to DL is still significant in all interferer scenarios at 10 dB and at 5-dB SIR, and in four out of seven scenarios at 0-dB SIR.

Spectrograms and example audio clips of clean speech, degraded speech, and enhanced speech are available from the authors' website.

VII. Conclusion

We have presented an enhancement method for speech degraded by nonstationary real-world interferer signals. Our method is based on learning speech and interferer signal models, and achieves significant improvements over geometric spectral subtraction and codebook-based filtering, both for light and considerable degradation of the clean speech signal.

We model the speech signal class and the interferer signal classes using learned dictionaries. An observation of degraded speech is enhanced by coding it in a composite dictionary, followed by separating the contributions to the observed mixture into a speech and an interferer contribution, based on the sparse coding weights. For this purpose we introduced LARC, a coding algorithm based on least angle regression, which employs a residual coherence threshold as the sparsity parameter. In contrast to specifying a coding cardinality or a residual norm value, it is not necessary to adapt the residual coherence threshold to the data on a frame by frame basis. Controlling the coding sparsity enables a tradeoff between source distortion and source confusion, i.e., controlling the amount of speech degradation versus interferer intrusion in the enhanced speech signal.

For high-dimensional feature spaces, the trained dictionary matrices contain millions of entries. The use of parametric functions for the dictionary atoms could reduce the size of the dictionary to a small number of parameters per atom [2]. This approach promises the possibility to handle larger dictionaries due to fewer model parameters, as well as enabling more efficient coherence computations if the parametric form has special structure, such as the Gabor atoms for which the coherence can be computed efficiently using the fast Fourier transform. Furthermore, the signal models could be extended to include a more descriptive prior probability distribution over the coding coefficients beyond independent sparsity priors, which would reduce the source confusion risk during enhancement.

Acknowledgment

The authors would like to thank P. Loizou for publishing the Matlab implementations of geometric spectral subtraction and the fwSegSNR and PESQ objective measures. Furthermore, the authors would like to thank the reviewers for their valuable comments and suggestions to improve this manuscript.

References

[1] M. Aharon, M. Elad, A. Bruckstein, and Y. Katz, "K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[2] M. Ataee, H. Zayyani, M. Babaie-Zadeh, and C. Jutten, "Parametric dictionary learning using steepest descent," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 1978–1981.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[4] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[5] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, pp. 2421–2424, 2006.

[6] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constr. Approx.*, vol. 13, pp. 57–98, 1997.

[7] G. M. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions with matching pursuits," *SPIE J. Opt. Eng.*, vol. 33, pp. 2183–2191, 1994.

[8] D. Donoho and M. Elad, "On the stability of the basis pursuit in the presence of noise," *Signal Process.*, vol. 86, no. 3, pp. 511–532, 2006.

[9] D. Donoho and I. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[10] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov. 2001.

[11] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, pp. 407–499, 2004.

[12] J. Eggert and E. Korner, "Sparse coding and NMF," in *Proc. IEEE Int. Conf. Neural Netw.*, 2004, pp. 2529–2533.

[13] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-qunatized phase-vocoder representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. 5, pp. 957–960.

[14] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[15] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 825–834, May 2008.

[16] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

[17] G. Kim and P. Loizou, "Improving speech intelligibility in noise using environment-optimized algorithms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2080–2090, Nov. 2010.

[18] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[19] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," *Adv. Neural Inf. Process. Syst.*, vol. 19, pp. 801–808, 2007.

[20] P. C. Loizou, *Speech Enhancement: Theory and Practice*. New York: Taylor & Francis, 2007.

[21] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Commun.*, vol. 50, pp. 453–466, 2008.

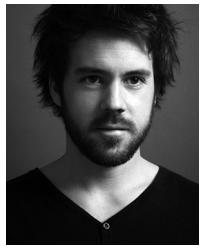
[22] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, pp. 3387–3405, 2009.

[23] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, 2010.

[24] S. Mallat, *A Wavelet Tour of Signal Processing – The Sparse Way*. New York: Academic, 2009.
[25] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[26] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[27] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1153–1161, Oct. 1986.
[28] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2210–2219, May 2008.
[29] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.
[30] S. Roweis, "One microphone source separation," *Adv. Neural Inf. Process. Syst.*, pp. 793–799, 2001.
[31] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010.
[32] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit. technical report," Technion, Haifa, 2008.
[33] M. Schmidt and J. Larsen, "Reduction of non-stationary noise using a non-negative latent variable decomposition," in *Proc. IEEE Workshop Mach. Learn. Signal Process.*, 2008, pp. 486–491.
[34] C. Sigg, T. Dikk, and J. Buhmann, "Speech enhancement with sparse coding in learned dictionaries," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4758–4761.
[35] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Independent Component Analysis and Signal Separation*. New York: Springer, 2007, pp. 414–421.
[36] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006.
[37] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.
[38] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Statist. Soc.*, vol. 58, pp. 267–288, 1996.
[39] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1978, pp. 586–590.
[40] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, *The Noisex-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. Technical Report*. Malvern, U.K.: DRA Speech Res. Unit, 1992.
[41] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4029–4032.

Christian D. Sigg (S'07–M'11) received the Dipl. Ing. degree in computer science and the Dr. sc. degree from ETH Zurich, Zurich, Switzerland, in 2004 and 2011, respectively.

From October 2005 to November 2011, he was with the Machine Learning Laboratory, ETH Zurich, first as a Ph.D. student and then as a Postdoctoral Researcher. In December 2011, he joined the Data Management Group at the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss), Zurich. His areas of interest include signal processing based on sparse coding principles and model inference for automated data quality control.



Tomas Dikk received the M.Sc. degree in computer science from ETH Zurich, Zurich, Switzerland, in 2008.

Since March 2008, he has been a Research Associate in the Machine Learning Laboratory at ETH Zurich. His research interests include speech enhancement based on sparse coding methods and audio-visual sensory fusion.

Joachim M. Buhmann (M'90–SM'06) received the Ph.D. degree in theoretical physics from the Technical University of Munich, Munich, Germany, in 1988.

He is a Professor for information science and engineering at the Computer Science Department, Swiss Federal Institute of Technology Zurich (ETH), Zurich, Switzerland. He has held postdoctoral and research faculty positions at the University of Southern California, Los Angeles, and the Lawrence Livermore National Laboratory, Livermore, CA, from 1988 to 1992. Until October 2003, he headed the Research Group on Pattern Recognition, Computer Vision, and Bioinformatics in the Computer Science Department, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany. In October 2003, he joined ETH Zurich. His current research interests cover machine learning, statistical learning theory and its relations to information theory as well as applications of machine learning to challenging data analysis questions. The machine learning applications range from image understanding and medical image analysis, to signal processing, bioinformatics, and computational biology. Special emphasis is devoted to model selection questions for the analysis of large-scale heterogeneous data sets.

Dr. Buhmann has served as an associate editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He is currently the president of the German pattern recognition society DAGM.