SPEECH ENHANCEMENT WITH SPARSE CODING IN LEARNED DICTIONARIES

Christian D. Sigg, Tomas Dikk and Joachim M. Buhmann

Department of Computer Science, ETH Zurich, Switzerland
{chrsigg,tdikk,jbuhmann}@inf.ethz.ch

ABSTRACT

The enhancement of speech degraded by non-stationary interferers is a highly relevant and difficult task of many signal processing applications. We present a monaural speech enhancement method based on sparse coding of noisy speech signals in a composite dictionary, consisting of the concatenation of a speech and interferer dictionary, both being possibly over-complete. The speech dictionary is learned off-line on a training corpus, while an environment specific interferer dictionary is learned on-line during speech pauses. Our approach optimizes the trade-off between source distortion and source confusion, and thus achieves significant improvements on objective quality measures like *cepstral distance*, in the speaker dependent and independent case, in several real-world environments and at low signal-to-noise ratios. Our enhancement method outperforms state-of-the-art methods like *multi-band spectral subtraction* and approaches based on vector quantization.

Index Terms— Speech Enhancement, Dictionary Learning, Sparse Coding, Source Separation.

1. INTRODUCTION

Enhancing speech degraded by interferers is an important task for many signal processing applications, including hearing aids, mobile communications and speech recognition [1]. The difficulty arises from the nature of real-world interferers that are often non-stationary and potentially speech-like, thereby inducing a significant and variable spectral overlap between speech and interferer. Furthermore, the class of possible interferer signals shows high variability, and interferers are typically a superposition of several sources, requiring a comprehensive interferer model to be prohibitively complex.

The goal of speech enhancement is to improve both the intelligibility and the quality of speech, by attenuating the interferer without substantially degrading the speech. As a substitute to perform subjective listening tests, objective measures like *cepstral distance* [1] quantify quality improvement by comparing the (unobserved) clean speech with the noisy speech and the enhanced speech in a perceptually meaningful way.

We consider the setting of a one-to-one conversation in a natural environment, recorded by a single microphone. This setup results in a linear additive mixture of clean speech and interferer. The clean speech is not directly observable, the interferer signal however is observed during speech pauses. Therefore, we learn a speech model on a training corpus. This approach is justified because speech has limited variability, and a pre-trained model remains largely valid during enhancement. The contrary is true for the interferer, for which learning and adaptation is hence performed during every speech pause, resulting in a model that is specific to the current environment.

What follows is a high-level overview of our method, which is described in detail in Section 2. Our enhancer is implemented in the short-time Fourier transform (STFT) magnitude domain. Assuming that the phase of the interferer can be approximated with the phase of the mixture (common in the derivation of spectral subtraction algorithms [1]), linear additivity holds in the STFT magnitude domain, too. A possibly over-complete dictionary of atoms is trained for both speech and interferer magnitudes (Section 2.1), which are then concatenated into a composite dictionary. In the enhancement step (Section 2.2), an observation of noisy speech is sparsely coded in the composite dictionary. As a result, the mixture of speech and interferer is explained by a sum of a linear combination of atoms from the speech dictionary and a linear combination of atoms from the interferer dictionary. The clean speech magnitude is estimated by disregarding the contribution from the interferer dictionary, preserving only the linear combination of speech dictionary atoms (analogously for the interferer). Finally, a Wiener-like filter (Section 2.2) is constructed from the estimated magnitudes and applied to the mixture magnitude, to obtain an estimate of the clean speech magnitude. This estimate is combined with the phase of the mixture to re-synthesize the time domain signal.

As will be explained in Section 2.2, speech and interferer magnitude estimation errors result from two different and competing effects. A too sparse coding of the speech induces an approximation error, which we denote by *source distortion*. A too dense coding avoids source distortion, but causes *source confusion*, by explaining some of the speech magnitude using interferer atoms (analogously for the interferer, for both effects). A vector quantization (VQ) based

This work was in part funded by CTI grant 8539.2;2 ESPP-ES.

enhancer explains the mixture magnitude using at most one atom from the speech dictionary and one atom from the interferer dictionary. This restriction introduces a significant source distortion and thereby substantial magnitude estimation error. Results of Section 4 demonstrate that a better trade-off between source distortion and source confusion is achieved with a linear combination of several atoms per dictionary, explaining the superior enhancement performance of our method.

2. METHOD

We consider a signal $\mathbf{x} \in \mathbb{R}^D$ and a *dictionary* $\mathbf{D} = \left[\mathbf{d}_{(1)} \cdots \mathbf{d}_{(L)}\right] \in \mathbb{R}^{D \times L}$ consisting of L unit-norm *atoms*, $\|\mathbf{d}_{(l)}\|_2$ =1, $l = 1, \ldots, L$. A *sparse coding* $\mathbf{c} \in \mathbb{R}^L$ of signal $\mathbf{x}$ in dictionary $\mathbf{D}$ defines a sparse linear combination of $K \ll L$ atoms, such that the approximation error $\|\mathbf{x} - \mathbf{Dc}\|_2$ is "sufficiently small". The observation that speech and other structured signals can be well approximated by few atoms of a suitably trained dictionary [2] lies at the core of our enhancement algorithm.

2.1. Dictionary Learning

Dictionary learning adapts an initial dictionary to a specific signal class (e.g. speech). It is the generalization of codebook learning for VQ [3]. Instead of representing a signal by a single codebook vector, it is represented by a linear combination of dictionary atoms. *Learning* the dictionary is crucial for successful enhancement, where speech must have a sparse representation in the speech dictionary, but not in the interferer dictionary (i.e. the dictionaries must have low *mutual coherence*, see Section 2.2). Constructive dictionaries that are not signal class specific do not satisfy this requirement.

Dictionary learning is the matrix factorization of a data matrix $\mathbf{X} = \left[\mathbf{x}_{(1)} \cdots \mathbf{x}_{(N)}\right] \in \mathbb{R}^{D \times N}$ into a dictionary $\mathbf{D}$ and a coding $\mathbf{C} = \left[\mathbf{c}_{(1)} \cdots \mathbf{c}_{(N)}\right] \in \mathbb{R}^{L \times N}$, given by

$$\arg\min_{\mathbf{D},\mathbf{C}} \|\mathbf{X} - \mathbf{D} \cdot \mathbf{C}\|_F^2 \,, \tag{1}$$

subject to a sparsity constraint on $\mathbf{C}$ and the unit norm constraint on $\mathbf{D}$. $\|\cdot\|_F$ denotes the Frobenius norm.

Matrix factorization is a difficult problem, since the joint optimization of $\mathbf{D}$ and $\mathbf{C}$ is non-convex. Iterative solvers yield locally optimal solutions, by alternating between optimizing the coding and the dictionary. We use the kSVD algorithm of Aharon *et al.* [3], implemented in Matlab by kSVDBox[1]. On-line dictionary learning algorithms also exist [4]. The two steps of kSVD are as follows.

Coding update. For each column $\mathbf{c}_{(n)}$, $n = 1, \ldots N$, perform *orthogonal matching pursuit (OMP)* regression with approximation parameter σ,

$$\arg\min \left\|\mathbf{c}_{(n)}\right\|_0 \tag{2}$$

$$\text{s.t.} \quad \left\|\mathbf{x}_{(n)} - \mathbf{Dc}_{(n)}\right\|_2 \leq \sigma. \tag{3}$$

Dictionary update. For each column $\mathbf{d}_{(l)}$, $l = 1, \ldots, L$, separate the contribution of atom $\mathbf{d}_{(l)}$ to the residual norm

$$\|\mathbf{X} - \mathbf{D} \cdot \mathbf{C}\|_F^2 = \left\|\mathbf{X} - \sum_{j=1}^{L} \mathbf{d}_{(j)} \mathbf{c}^{[j]}\right\|_F^2 \tag{4}$$

$$= \left\|\left(\mathbf{X} - \sum_{j \neq l} \mathbf{d}_{(j)} \mathbf{c}^{[j]}\right) - \mathbf{d}_{(l)} \mathbf{c}^{[l]}\right\|_F^2$$

$$= \left\|\mathbf{R}^{(l)} - \mathbf{d}_{(l)} \mathbf{c}^{[l]}\right\|_F^2 \,, \tag{5}$$

where $\mathbf{c}^{[j]}$ is the j-th row of $\mathbf{C}$.

The residual norm is minimized w.r.t. $\mathbf{d}_{(l)}$ and $\mathbf{c}^{[l]}$ using the SVD. Define $\tilde{\mathbf{R}}^{(l)}$ as the set of columns of $\mathbf{R}^{(l)}$ indexed by $\left\{n | \mathbf{c}^{[l]}(n) \neq 0, 1 \leq n \leq N\right\}$ where atom $\mathbf{d}_{(l)}$ was involved in the coding. Compute the SVD

$$\tilde{\mathbf{R}}^{(l)} = \mathbf{U \Sigma V}^\top, \tag{6}$$

and update $\mathbf{d}_{(l)}$ as the first column of $\mathbf{U}$, and $\mathbf{c}^{[l]}$ as $\Sigma_{1,1}$ times the first row of $\mathbf{V}^\top$.

2.2. Enhancement

As discussed in the introduction, we assume that the observed noisy speech magnitude is the linear additive *mixture* $\mathbf{x} = \mathbf{s} + \mathbf{i}$ of clean speech magnitude $\mathbf{s} \in \mathbb{R}^D$ and interferer magnitude $\mathbf{i} \in \mathbb{R}^D$. The goal of the enhancement step is to obtain an estimate $\hat{\text{s}}$ of clean speech and an estimate $\hat{\mathbf{i}}$ of the interferer, given $\mathbf{x}$, a speech dictionary $\mathbf{D}_s \in \mathbb{R}^{D \times L_s}$ and an interferer dictionary $\mathbf{D}_i \in \mathbb{R}^{D \times L_i}$. For the formal analysis, we distinguish between *unstructured* and *structured* interferers (e.g. Gaussian white noise and background music, respectively), and make use of two results from sparse coding theory to enhance noisy speech in the presence of both. Due to space constraints, we omit detailed references, they can be found in chapters 11 and 12 of [2].

Unstructured interferer. An interferer is unstructured if it cannot be sparsely coded in any fixed dictionary, in particular not in $\mathbf{D}_s$. The coding of a mixture $\mathbf{x}$ in $\mathbf{D}_s$ therefore distributes the energy of the unstructured interferer contribution over all atoms of $\mathbf{D}_s$. OMP coding (eq. 2) of a mixture of speech and Gaussian white noise, with σ set to the noise standard deviation, correctly recovers the atoms in $\mathbf{D}_s$ which were responsible for the speech contribution to the mixture. This provides an accurate estimate $\hat{\text{s}}$ of the clean speech.

Structured interferer. An interferer is structured if it can be sparsely coded in a suitable dictionary $\mathbf{D}_i$. Define the composite dictionary $\mathbf{D} = [\mathbf{D}_s \, \mathbf{D}_i]$ as the concatenation of the

[1] http://www.cs.technion.ac.il/~ronrubin/software.html

speech and interferer dictionaries. The mixture $\mathbf{x}$ is coded in $\mathbf{D}$ using LASSO regression

$$\arg\min_{\mathbf{c}} ||\mathbf{x} - \mathbf{D}\mathbf{c}||_2 \tag{7}$$

$$= \quad \arg\min_{\mathbf{c}^s, \mathbf{c}^i} \left\| \mathbf{x} - [\mathbf{D}_s \, \mathbf{D}_i] \left[\begin{array}{c} \mathbf{c}^s \\ \mathbf{c}^i \end{array} \right] \right\|_2 \tag{8}$$

$$\text{subject to} \quad \frac{||\mathbf{c}||_1}{||\mathbf{x}||_2} \leq \theta, \tag{9}$$

with aptly chosen sparsity parameter θ (division by $||\mathbf{x}||_2$ normalizes variable signal gain), where $\mathbf{c}^s \in \mathbb{R}^{L_s}$ and $\mathbf{c}^i \in \mathbb{R}^{L_i}$. We estimate the clean speech as $\hat{\mathbf{s}} = \mathbf{D}_s\mathbf{c}^s$, and the interferer as $\hat{\mathbf{i}} = \mathbf{D}_i\mathbf{c}^i$. Lasso regularizes $\mathbf{c}$ by the penalty $\|\mathbf{c}\|_1$, which produces more stable speech and interferer estimates than OMP regression with the penalty $\|\mathbf{c}\|_0$. OMP codings can have large weights of opposite sign, which become apparent when separating the coding $\mathbf{c}$ into $\mathbf{c}^s$ and $\mathbf{c}^i$.

The estimation errors $\|\mathbf{s} - \hat{\mathbf{s}}\|_2$ and $\|\mathbf{i} - \hat{\mathbf{i}}\|_2$ are small if $\mathbf{s}$ and $\mathbf{i}$ in fact can be sparsely coded in their respective dictionaries, and if $\mathbf{D}_s$ and $\mathbf{D}_i$ have low *mutual coherence*

$$\mu(\mathbf{D}_s, \mathbf{D}_i) = \max_{1\leq p\leq L_s, 1\leq q\leq L_i} \left| \mathbf{d}^{s\top}_{(p)} \mathbf{d}^i_{(q)} \right|, \tag{10}$$

where $\mathbf{d}^s_{(p)}$ is the p-th atom of $\mathbf{D}_s$. If an exact speech coding $\tilde{\mathbf{c}}^s$ exists, i.e. $\mathbf{s} = \mathbf{D}_s\tilde{\mathbf{c}}^s$ (analogously for the interferer), and the *exact recovery condition (ERC)*

$$\|\tilde{\mathbf{c}}^s\|_0 + \|\tilde{\mathbf{c}}^i\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}_s, \mathbf{D}_i)}\right) \tag{11}$$

is fulfilled, the LASSO coding $\mathbf{c}$ explains $\mathbf{s}$ using only atoms from $\mathbf{D}_s$, and $\mathbf{i}$ using only atoms from $\mathbf{D}_i$.

Because real-world interferers might not allow for a sufficiently sparse coding, and because they can have speech like properties increasing the mutual coherence, the ERC might not be satisfied. As a consequence, a too dense coding of $\mathbf{x}$ in $\mathbf{D}$ introduces *source confusion*, explaining some of the energy in $\mathbf{s}$ using atoms from $\mathbf{D}_i$ (analogously for $\mathbf{i}$). On the other hand, a too sparse coding of $\mathbf{x}$ in $\mathbf{D}$, although avoiding source confusion, increases *source distortion* by coding $\mathbf{s}$ with too few atoms of $\mathbf{D}_s$ (analogously for $\mathbf{i}$). We have observed empirically that both effects contribute to estimation errors $\|\mathbf{s} - \hat{\mathbf{s}}\|_2$ and $\|\mathbf{i} - \hat{\mathbf{i}}\|_2$, and that the minimum is attained by a trade-off between both effects.

Varying the sparsity of $\mathbf{c}$ (with LASSO parameter θ) controls the trade-off between source confusion and source distortion. By choosing the optimal θ^*, our method lowers the source distortion significantly more than increasing the estimation error due to source confusion. In contrast, the VQ based enhancer selects only one atom per dictionary, resulting in small source confusion, but significant source distortion.

Wiener-like filtering. A filter constructed from $\hat{\mathbf{s}}$ and $\hat{\mathbf{i}}$ is applied to the mixture $\mathbf{x}$, to obtain the final clean speech magnitude estimate

$$\mathbf{s}_w = \hat{\mathbf{s}} \oslash (\hat{\mathbf{s}} + \hat{\mathbf{i}}) \otimes \mathbf{x}, \tag{12}$$

where $\oslash$ and $\otimes$ denote element-wise division and multiplication. Note that if $\hat{\mathbf{s}}$, $\hat{\mathbf{i}}$ and $\mathbf{x}$ were power spectra, (12) would correspond to a Wiener filter. Finally, $\mathbf{s}_w$ is combined with the mixture phase to re-synthesize the time-domain signal.

3. RELATED WORK

We compare our method to two established speech enhancement approaches, VQ based enhancement and multi-band spectral subtraction. Srinivasan *et al.* [5] pre-train short-term linear prediction codebooks for speech and interferer signals. They avoid the full complexity of considering all pairs of an element from the speech codebook with an element of the noise codebook by an iterative element selection strategy. For our evaluation, we implemented a VQ based enhancer that shares the same pipeline as our method, but chooses one atom from the speech dictionary and one atom from the interferer dictionary only, using a greedy selection strategy.

We also compare our method to multi-band spectral subtraction, using the implementation of [1]. This state-of-the-art method achieves equal or better subjective listener ratings [1] than many other approaches (e.g. subspace methods and statistical model based methods). Spectral subtraction only models the interferer, by averaging the interferer magnitude spectrum during a speech pause. However, estimating only average interferer magnitude limits enhancement performance, because the interferer contribution to the mixture at some point in time can deviate significantly from the average.

Our work has conceptual similarities to the single-channel speaker separation approach of Schmidt and Olsson [6], where the authors used sparse non-negative matrix factorization (sNMF) to train speaker dependent dictionaries, and separated an anechoic mixture of two speakers by sNMF coding in the concatenated dictionary. We show that the same fundamental idea can be successfully extended to speech enhancement, and complement it by providing insight into the conditions for enhancement in real-world environments, where the theoretical guarantees of the ERC don't hold.

4. EVALUATION

We predict in Section 2.2 that our method achieves a better source distortion and confusion trade-off than VQ based enhancement. This translates into significantly higher quality improvements, quantified by the cepstral distance measure on validation data. In addition, we provide baseline results of the multi-band spectral subtraction enhancer. Example spectrograms, audio clips and results for additional objective measures are available on the web[2].

As speech data, we use recordings from the Grid corpus[3]. As non-stationary interferers, four recordings taken in

[2] http://www.inf.ethz.ch/personal/chrsigg/icassp2010

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

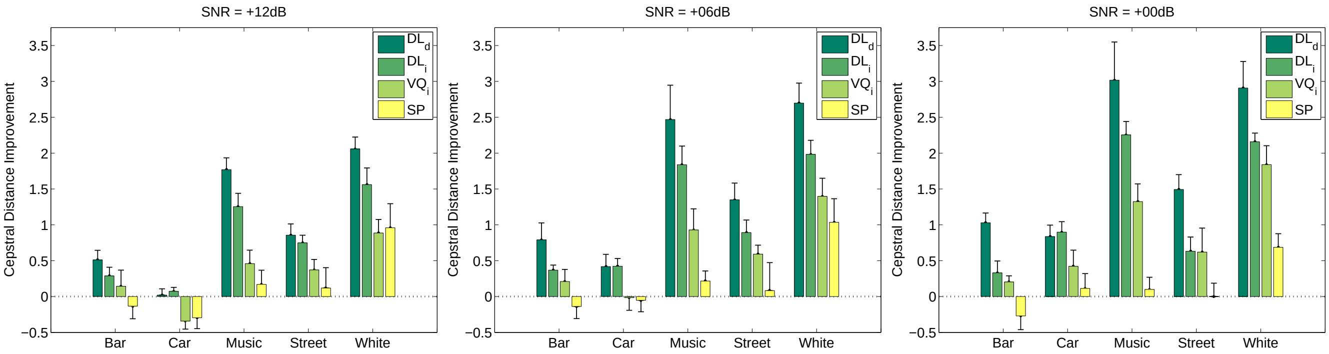

Fig. 1. Objective quality improvements, measured by the difference of cepstral distance of noisy to clean speech, and the cepstral distance of enhanced to clean speech. We compare our method in the speaker dependent (DL_d) and independent case (DL_i) to VQ based speaker independent enhancement (VQ_i) and spectral subtraction (SP). Filled bars denote median cepstral distance improvements, error bars denote 25% quantiles for negative and 75% quantiles for positive improvements.

real-world environments are used: speech babble and clatter noise in a bar, engine and tire noise in a car, classical piano music replayed indoors, street traffic noise, and white noise as a maximally non-sparse interferer. The data is randomly split into train, test and validation sets, using a 9:3:3 ratio. The speaker dependent experiment uses one male speaker, the speaker independent experiment uses 30 speakers of both genders. The time-domain signals are transformed into STFT frames ($D = 129$). Mixtures are synthetically generated by adding clean speech and interferer at various SNRs, since objective measures require access to the clean speech signal.

Dictionaries are trained using kSVD, initialized with atoms sampled uniformly on the unit sphere. The optimal parameter σ^* (eq. 3) for each dictionary size L is determined on test data. For the enhancement, the optimal combination of dictionary sizes L_s^* and L_i^*, as well as the optimal sparsity parameter θ^* at a given SNR, is again determined on test data.

Enhancement performance is measured by the difference of cepstral distance [1] of noisy to clean speech, and the cepstral distance of enhanced to clean speech. A positive improvement implies a reduction of cepstral distance, and a negative improvement implies that artifacts introduced by the enhancer degrade the noisy speech even further. Figure 1 reports the performance of our method in the speaker dependent (DL_d) and speaker independent case (DL_i), compared to VQ based speaker independent enhancement (VQ_i) and spectral subtraction (SP). Both DL_d and DL_i significantly outperform VQ_i and SP at +12dB and +6dB SNR. At 0dB SNR, the median quality improvement of DL_i compared to VQ_i is less significant for street traffic and white noise, as source confusions become increasingly likely.

5. CONCLUSION

We presented a speech enhancement method based on sparse coding in learned dictionaries. The method integrates key results from the dictionary learning and sparse coding literature, to provide effective enhancement of speech in the presence of real-world non-stationary and potentially speech-like interferers. Our method explicitly optimizes the source distortion and source confusion trade-off, which translates into significantly higher quality improvements than highly competitive VQ and state-of-the-art multi-band spectral subtraction enhancers.

Currently, the speech and interferer contributions to the signal mixture are estimated by maximum-likelihood. We plan to develop a fully Bayesian framework by introducing suitable prior distributions on speech and interferer codings. Furthermore, incorporating the visible dynamics of speech production could provide valuable side information for further improving performance in very low SNR situations.

6. REFERENCES

[1] Philipos C. Loizou, *Speech Enhancement: Theory and Practice*, Taylor and Francis, 2007.

[2] Stephane Mallat, *A Wavelet Tour of Signal Processing - The Sparse Way*, Academic Press, 2009.

[3] Michal Aharon, Michael Elad, Alfred Bruckstein, and Yana Katz, “K-svd: An algorithm for designing of overcomplete dictionaries for sparse representation,” *IEEE Trans. Signal Processing*, 2006.

[4] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, “Online dictionary learning for sparse coding,” in *Proc. 26th ICML*, 2009.

[5] S. Srinivasan, J. Samuelsson, and WB Kleijn, “Codebook driven short-term predictor parameter estimation for speech enhancement,” *IEEE Trans. ASLP*, 2006.

[6] Mikkel N. Schmidt and Rasmus K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in *Proc. 9th ICSLP*, 2006.