SPEECH ENHANCEMENT SYSTEM USING FISCHER DISCRIMINATIVE DICTIONARY LEARNING FDDL

Dima SHAHEEN[1], Oumayma AL-DAKKAK[2], Mohiedin WIANAKH[3]

Speech enhancement is one of the challenging tasks in signal processing, especially in the case of non-stationary speech-like noise. In this paper we propose a new supervised speech enhancement system that uses Fischer Discriminative Dictionary Learning (FDDL) algorithm to model both speech and noise amplitude spectrum, where the cost function accounts for both "source confusion" and "source distortion" errors. In the enhancement stage, we use sparse coding on the learnt dictionary to find an estimate for both the clean speech and noise amplitude spectrum. In the final stage, a Wiener filter is used to refine the clean speech estimate. Experiments on NOIZEUS dataset using two objective speech enhancement measures: frequency-weighted segmental SNR and Perceptual Evaluation of Speech Quality (PESQ) demonstrate that FDDL outperforms other tested dictionary learning algorithms in the presence of considerable noise (0 dB) for all studied noise types, and in the presence of structured non-stationary noise (ex. car and train noise) for all noise levels.

Keywords: speech enhancement, supervised dictionary learning, generative dictionary learning (GDL), sparse coding

1 Introduction

Speech enhancement is the task of extracting clean speech signal from a noisy mixture. It is a challenging task as it is hard to remove noise efficiently without distorting the estimated clean speech signal. The main goal for speech enhancement algorithms is two folds: to enhance "speech quality", which refers to the ease for humans to listen to the enhanced speech for a long time, and to ameliorate "speech intelligibility", which refers to the reduction of word error rate when using the enhanced speech.

The problem we are tackling in this paper is the single channel speech enhancement that aims to reconstruct the clean speech signal $s(n)$, based on the

[1] PhD student, Dept. of telecommunication, Higher Institute for Applied Science and Technology HIAST, Damascus, Syria, e-mail: dima.shaheen@hiast.edu.sy

[2] Prof., Dept. of telecommunication, Higher Institute for Applied Science and Technology HIAST, Damascus, Syria, email: oumayma.dakkak@hiast.edu.sy

[3] Prof., Dept. of telecommunication, Higher Institute for Applied Science and Technology HIAST, Damascus, Syria, email: mohiedin.wainakh@hiast.edu.sy

received signal (n), which is an additive mixture of the two unknown signals: the clean speech and a noise signal $i(n)$:

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

Traditional speech enhancement methods like Spectral Subtraction (SS) [1,2], Wiener filtering [3], statistical-model based methods [4], and SubSpace Approach (SSA) [5,6] perform well in the case of white noise, but have limited performance in the case of non-stationary speech-like noise. SS is based on estimating the noise power spectrum and subtracting it from the noisy power spectrum. The main issue with SS is the generation of isolated peaks in the estimated clean speech spectrum, which is referred to as musical noise. All these methods are unsupervised; they do not use any prior information about the noise or the speech. Recently, new supervised methods [7-12,16] based on incorporating prior information to build a model for both clean speech and noise signals using training samples have been proposed. These methods achieve better results than non-supervised ones.

Motivated by the great success of the sparsity based signal model achieved in many signal processing tasks, and notably image denoising [18], Sigg [13] proposed using the approximate K-Singular Value Decomposition (K-SVD) [19,20] dictionary learning to model the amplitude spectrum of the clean speech and the noise separately, and then concatenating both dictionaries in one to perform speech enhancement.

Zhao et al. [14] proposed using the same K-SVD with a non-negative constraint at the sparse coding stage to learn a dictionary that models the Power Spectral Density (PSD) of the clean speech, and used Least Angle Regression algorithm (LARS) [24] to find the sparse code of the noisy speech on the learned dictionary. Then clean speech PSD is estimated using the multiplication of the sparse code with the dictionary. Luo et al. [15] proposed a complementary joint sparse representation, where two mixture dictionaries: “mixture and speech” and “mixture and noise” are being added to the Generative Dictionary Learning (GDL) problem formulation. Sparse codes of the clean speech are forced to represent the noisy mixture on the mixture and clean speech sub-dictionary, while sparse codes of the noise are forced to represent the noisy mixture on the “mixture and noise” sub-dictionary. Though this joint sparse representation alleviates to some extent the problem of source confusion, but it has high complexity due to the need of learning four sub-dictionaries instead of two.

In the previous studies only “signal approximation” is considered in the cost function when learning the representative dictionaries, while “*source confusion*” between speech and noisy sub-dictionaries are not taken into account in the Dictionary Learning (DL) process. Source confusion means that part of the noise that is coherent to the clean speech will have sparse representation over the clean speech dictionary (noise confusion), and part of the clean speech will have sparse representation over the noise dictionary (speech confusion), and thus in the enhancement stage, noise residual corresponding to "noise confusion" might still exist in the estimated clean speech, which will also suffer from extra distortion from the original clean speech due to the fact that part of it corresponding to "speech confusion" will be omitted as it will be considered as noise.

In this paper we propose a new speech enhancement system that uses Fischer Discriminative Dictionary Learning (FDDL) [21] algorithm to model both speech and noise jointly in a way that minimizes source confusion error.

The paper is organized as follows. In Section 2 we provide a review of the main problems: dictionary learning algorithms and speech enhancement using sparse coding. In Section 3 the proposed speech enhancement system is being described. In Section 4 we present the conducted experiments and results. And in Section 5 we summarize and conclude the paper.

2 Literature Review

2.1 Dictionary Learning

Sparsity based signal model approximates a signal by a linear combination of few basic signals out of a larger collection of signals that form what is called the dictionary.

In the classical dictionary learning problem, we seek a matrix $\mathbf{D} \in \mathbb{R}^{N\times K}$, whose K columns are the basic signals that can represent the training signals $\mathbf{y}_{\mathrm{i}} \in \mathbb{R}^{N}$ *sparsely* as close as possible:

$$\min_{\mathbf{D},\mathbf{X}} \sum_{i=1}^{n} (\|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 \quad s.t. \quad \forall i, \|\mathbf{x}_i\|_0 \le k \tag{2}$$

Where *k* is the maximum number of non-zero elements in the sparse code $\mathbf{x}_i \in \mathbb{R}^K$, n is the number of training samples, $\|\mathbf{x}_i\|_0$ is ℓ_0 pseudo norm, which represents the number of non-zeros in $\mathbf{x}_i$, $\mathbf{X}$ is the matrix composed of all the sparse code vectors $\mathbf{x}_i$.

This optimization problem is nonconvex when both $\mathbf{D}$ and $\mathbf{X}$ are unknown, however, it becomes convex if one of $\mathbf{D}$ or $\mathbf{X}$ is fixed, that is why it is generally solved iteratively by fixing the dictionary $\mathbf{D}$ and updating the sparse codes $\mathbf{X}$, and then fixing $\mathbf{X}$ and updating $\mathbf{D}$.

How to update the dictionary atoms is the key difference between dictionary learning algorithms. Some dictionary learning methods update, in each iteration, the whole set atoms. This is the case of one of the early and simple dictionary learning MOD (Method of Optimal Direction) [17], which updates the whole dictionary using the closed form of the Mean Squared Error (MSE) estimator .

Other dictionary learning algorithms update the dictionary atoms successively one by one, like the case of the very famous and successful dictionary learning algorithm K-SVD [19]. In the sparse coding stage, K-SVD uses greedy Orthogonal Matching Pursuit (OMP) [23] to find the sparse code for each training sample. While in the dictionary update stage, for each dictionary atom, K-SVD selects only the training samples that use this atom. The update of the selected dictionary atom is done in a way that minimizes the restricted error (which is the only part of the total sparse representation error that is affected by the selected atom), and for this purpose K-SVD evaluates the Singular Value Decomposition (SVD) of the restricted error [19].

The cost function in (2) only measures the representation power of the Dictionary **D**. In the case of classification task, discriminative power of the sparse code **x** should be considered. This leads to a new trend in dictionary learning algorithms called “discriminative” or “supervised” dictionary learning in which the cost function reflects both the representation and classification error.

FDDL [21] uses labels information in both the dictionary update stage and the sparse coding stage. In FDDL, the sparse codes of the training samples are forced to have small within-class scatter but big between-class scatter. Also, each class-specific sub-dictionary is forced to have good reconstruction capability for the training samples from that class but poor reconstruction capability for other classes. Therefore, both the representation residual and the representation coefficients of the query sample are discriminative.

FDDL has been applied to image classification [21], face recognition [34], and facial expression recognition [35].

2.2 Speech Enhancement using Sparse Coding

Sigg [13] proposed a supervised speech enhancement method based on learning two dictionaries, one for clean speech and another for noise, according to the following formulations:

$$\min_{\mathbf{D}_s \mathbf{X}_s} \|\mathbf{Y}_s - \mathbf{D}_s\mathbf{X}_s\|_F^2 \quad s.t. \|\mathbf{X}_s\|_0 \le k_s \tag{3}$$

$$\min_{\mathbf{D}_n, \mathbf{X}_n} \|\mathbf{Y}_n - \mathbf{D}_n\mathbf{X}_n\|_F^2 \quad s.t. \|\mathbf{X}_n\|_0 \le k_n \tag{4}$$

Sigg proposed GDL to solve each of the previous problems. GDL is in fact, a variation of the approximate K-SVD [20], the only difference is at the sparse coding stage. Sigg proposed Least Angle Regression with Coherence Criterion (LARC) [13] for sparse coding, instead of the greedy OMP [23]. LARC is a variation of the LARS [24], where the coherence between the residual error and the dictionary is being used as stopping criterion instead of the l_2 norm of the residual error.

The problem with GDL is that the two sub-dictionaries $\mathbf{D}_s$ and $\mathbf{D}_n$ are being learnt independently (see equation (3) and (4)), and thus source confusion error is not being considered. FDDL algorithm takes into consideration this problem, and the objective is to learn a class specific dictionary that has low representation error for samples of the same class, and high representation error for samples of other classes (low confusion).

3 Speech Enchantement System using FDDL

The proposed speech enhancement system is depicted in Fig. 1. The system contains two stages: training and enhancement. In the training stage we learn the FDDL dictionary that models the amplitude spectrum of the training clean speech and noise samples. The amplitude of the Short Time Fourier coefficients STFT for the overlapping time frames of the clean speech and noise training signals are calculated after applying a Hamming window. The amplitude spectrum coefficients for all training frames are concatenated as columns to form $\mathbf{Y}_s$ and $\mathbf{Y}_n$, and fed to the FDDL algorithm that learns the clean speech sub-dictionary $\mathbf{D}_\mathrm{s}$, and the noise sub-dictionary $\mathbf{D}_\mathrm{n}$. These two sub-dictionaries are concatenated together to form the overall dictionary, that contains $2L$ coulmns.

At the enhancement phase, using LARC and the dictionary $\mathbf{D}$, the sparse codes $\mathbf{X}$ for the amplitude spectrum coefficients of the overlapping frames of the noisy signal are calculated. The columns of $\mathbf{X}$ are the sparse code vectors $(N$ is the number of FFT coefficients), which contain $2L$ coefficients; the first L ones $\mathbf{X}_\mathrm{s}$ that correspond to the sub-dictionary $\mathbf{D}_\mathrm{s}$ are separated from the last L ones $\mathbf{X}_\mathrm{n}$ that correspond to the sub-dictionary $\mathbf{D}_\mathrm{n}$. By multiplying $\mathbf{X}_\mathrm{s}$ by $\mathbf{D}_\mathrm{s}$, and multiplying $\mathbf{X}_\mathrm{n}$ by $\mathbf{D}_\mathrm{n}$ we get an initial estimation for the amplitude spectrum of the clean speech and noise signals respectively. These initial estimations are fed to Wiener filter to find the final clean speech amplitude spectrum estimation. Finally, we apply Inverse Fourier Transform to the estimated amplitude spectrum combined with the noisy phase spectrum to get the estimated clean speech.

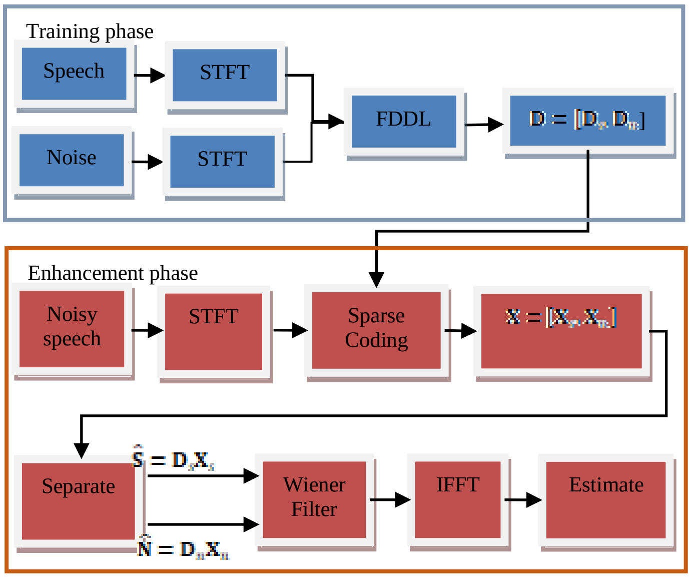

Fig. 1. The overall speech enhancement system

In the previous setting, FDDL [21] iterates between two steps. After initializing the dictionary $\mathbf{D}$, the first step is to find the sparse codes $\mathbf{X}$ of the training samples $\mathbf{Y}$ over the dictionary $\mathbf{D}$. This includes finding $\mathbf{X}_s \in \mathbb{R}^{K\times n_s}$ the sparse codes of the speech samples $\mathbf{Y}_s \in \mathbb{R}^{N\times n_s}$($n_s$ is the number of clean speech training frames, and $\mathbf{X}_n \in \mathbb{R}^{K\times n_n}$the sparse codes of the noise samples $\mathbf{Y}_n \in \mathbb{R}^{N\times n_n}$ (n_n is the number of noise training frames) (equations (5),(6)). In the second step, the sparse codes $\mathbf{X}$ (concatenation of $\mathbf{X}_s$ and $\mathbf{X}_s$) are fixed while the total dictionary $\mathbf{D}$ is updated. First, the sub-dictionary specific for representing noise samples $\mathbf{D}_n \in \mathbb{R}^{N\times L}$ is fixed while the sub-dictionary specific for representing clean speech samples $\mathbf{D}_s \in \mathbb{R}^{N\times L}$ is updated (equation (7)), then $\mathbf{D}_s$ is fixed while $\mathbf{D}_n$ is updated (equation (8)). $\mathbf{D}_s$ and $\mathbf{D}_n$ who has the same number of atoms L are concatenated to form the total dictionary $\mathbf{D}$, which has $K = 2L$ atoms.

1. Update $\mathbf{X}$ $(\mathbf{D} = [\mathbf{D}_s, \mathbf{D}_n]$ is fixed):

a. Update $\mathbf{X}_s$ $(\mathbf{X}_n, \mathbf{D}$ are fixed)

$$\mathbf{X}_s = \min_{\mathbf{X}_s} \|\mathbf{Y}_s - \mathbf{D}\mathbf{X}_s\|_F^2 + \|\mathbf{Y}_s - \mathbf{D}_s\mathbf{X}_s^s\|_F^2 + \|\mathbf{D}_n\mathbf{X}_s^n\|_F^2 + \lambda_1.\|\mathbf{X}_s\|_1 + \lambda_2.(\|\mathbf{X}_s - \mathbf{M}_s\|_F^2 - \|\mathbf{M}_s - \mathbf{M}\|_F^2 - \|\mathbf{M}_n - \mathbf{M}\|_F^2 + \eta.\|\mathbf{X}_s\|_F^2) \quad (5)$$

b. Update $\mathbf{X}_n$ $(\mathbf{X}_s, \mathbf{D}$ are fixed)

$$\mathbf{X}_n = \min_{\mathbf{X}_n} \|\mathbf{Y}_n - \mathbf{D}\mathbf{X}_n\|_F^2 + \|\mathbf{Y}_n - \mathbf{D}_s\mathbf{X}_n^n\|_F^2 + \|\mathbf{D}_s\mathbf{X}_n^s\|_F^2 + \lambda_1.\|\mathbf{X}_n\|_1 + \lambda_2.(\|\mathbf{X}_n - \mathbf{M}_n\|_F^2 - \|\mathbf{M}_n - \mathbf{M}\|_F^2 - \|\mathbf{M}_s - \mathbf{M}\|_F^2 + \eta.\|\mathbf{X}_n\|_F^2) \quad (6)$$

2. Update $\mathbf{D}$ $(\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_n]$ fixed)

a. Update $\mathbf{D}_s$ $(\mathbf{X}, \mathbf{D}_n$ are fixed):

$$\mathbf{D}_s = \min_{\mathbf{D}_s} \|(\mathbf{Y} - \mathbf{D}_n\mathbf{X}^n) - \mathbf{D}_s\mathbf{X}^s\|_F^2 + \|\mathbf{Y}_s - \mathbf{D}_s\mathbf{X}_s^s\|_F^2 + \|\mathbf{D}_s\mathbf{X}_n^s\|_F^2$$

$$s.t. \left\|\mathbf{d}_{s,l}\right\|_2 = 1 \, , l = 1..L \quad (7)$$

b. Update $\mathbf{D}_n$ ($\mathbf{X}, \mathbf{D}_s$ fixed):

$$\mathbf{D}_n = \min_{\mathbf{D}_n} \|(\mathbf{Y} - \mathbf{D}_s\mathbf{X}^s) - \mathbf{D}_n\mathbf{X}^n\|_F^2 + \|\mathbf{Y}_n - \mathbf{D}_n\mathbf{X}_n^n\|_F^2 + \|\mathbf{D}_n\mathbf{X}_s^n\|_F^2$$

$$s.t. \left\|\mathbf{d}_{n,l}\right\|_2 = 1, l = 1..L \quad (8)$$

In the previous equations: $\mathbf{X}_s^s \in \mathbb{R}^{L\times n_s}$ is matrix formed by the first L rows of $\mathbf{X}_s$, $\mathbf{X}_s^n \in \mathbb{R}^{L\times n_s}$ is the last L rows of $\mathbf{X}_s \in \mathbb{R}^{2L\times n_s}$. The same applies to $\mathbf{X}_n^s, \mathbf{X}_n^n \in \mathbb{R}^{L\times n_n}$ (first and last L rows of $\mathbf{X}_s$), and to $\mathbf{X}^s, \mathbf{X}^n \in \mathbb{R}^{L\times n}$ (first and last L rows of $\mathbf{X} \in \mathbb{R}^{2L\times n}$). $\mathbf{M}_s$, $\mathbf{M}_n$ and $\mathbf{M}$ are the mean vector matrices (by taking $\boldsymbol{m}_s$ the mean of $\mathbf{X}_s$, $\boldsymbol{m}_n$ the mean of $\mathbf{X}_n$, $\boldsymbol{m}$ the mean of $\mathbf{X}$, as all the columns vectors)[21]. $\|\mathbf{Y}_s - \mathbf{D}_s\mathbf{X}_s^s\|_F^2$ represents the speech distortion error, $\|\mathbf{D}_s\mathbf{X}_n^s\|_F^2$ represents the noise confusion error, $\|\mathbf{Y}_n - \mathbf{D}_n\mathbf{X}_n^n\|_F^2$ represents the noise distortion error, and $\|\mathbf{D}_n\mathbf{X}_s^n\|_F^2$ represents the speech confusion error. Thus we can see that FDDL tries to minimize both distortion and confusion errors.

Yang [21] showed that equation (5) and (6) can be solved using Fast Iterative Shrinkage and Thresholding Algorithm (FISTA) [25], and problems (7) and (8) can be solved using the algorithm described in [33]. Vu [22] proposed an efficient implementation for FDDL (E-FDDL) where he used the iterative (ADMM) Alternating Direction of Multipliers Method [32] for solving (7) and

(8).

4 Experiments

4.1 Noizeus Dataset

Noizeus [27] is a noisy database that contains 30 IEEE sentences produced by three male and three female speakers, with 5 different sentences per speaker. The sentences are corrupted by eight different real-world noises at different SNRs: (0,5,10,15) dB. The noise was taken from the AURORA database and includes suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise [27]. All speech and noise signals are sampled at 8 kHz.

As the database contains a small number of speakers, for the speaker independent case, we have divided the dataset into two sets: a training set containing three speakers and a testing set containing the left three other speakers. We have created 10 training/test sets through permutations of three speakers out of 6 and averaged the results. All the training sets contain male and female speakers.

To assess the performance of our proposed speech enhancement system we compared its performance in terms of two objective measures: Frequency Weighted Segmental SNR (fwSegSNR) and Perceptual Evaluation of Speech Quality (PESQ) against two other different DL algorithms: K-SVD and GDL.

FwSegSNR is the estimated mean frequency domain SNR over all the time frames, with a perceptually motivated frequency band weighting. FwSegSNR can be calculated through the following equation:

$$\frac{10}{Nw}\sum_{i=1}^{n}\sum_{b=1}^{B} w_b \log_{10}\frac{|S(b,i)|^2}{\left(|S(b,i)|-|\hat{S}(b,i)|\right)^2} \qquad (9)$$

Where $S(b,i)$ are the complex FFT coefficients of the clean speech, i is the frame index, b is the frequency component index, n is the total number of frames in the speech signal, B is the total number of frequency components, w_b is the corresponding frequency weighting, w is the sum of all the frequency weights, and $\hat{S}(b,i)$ are the estimated complex spectrum coefficients of the enhanced speech.

PESQ [28] is widely used international measure to assess speech quality through telephony network. Its derivation can be found in [28]. For both measure we have used the implementation provided by [29].

4.2 Experimental Results

We have used the same speech enhancement system depicted in Fig. 1, but with the different DL algorithms: FDDL, K-SVD and GDL. Different frame lengths were investigated; starting with 256 up to 1024 samples (from 32 ms to 128 ms) with 50% overlapping. We have found that longer frames always give better results, as this will increase the dimension of the feature space and thus results in lower coherence between the clean speech and noise sub-dictionaries, which means lower source confusion error. The frame length used in the reported results below is 128 ms, and the number of DFT coefficients is 1024, only half of them is kept because of symmetry. For LARC, the stopping residual coherence thresholds is set to µ=0.15 in GDL at the training stage, while it is set to µ=0.1 for sparse coding at the enhancement stage, as has been noted by Sigg in [13]. For FDDL implementation, we have used the efficient implementation provided by Vu [22] denoted by E-FDDL, where two parameters are to be tuned: λ_1, and λ_2 (equation (6)). Following the same procedure in [21], we have tried different values spanning the set {0, 0.001, 0.005, 0.01, 0.05} for both parameters, on a validation test group, and found that $\lambda_1 = 0.05, \lambda_2 = 0.01$ give the best performance in terms of both performance measures. Experiments were conducted using Matlab 2015Ra on a laptop with 3.16 GHz Intel core i5 processor and 4 GB RAM. The reported results listed below are for the case where the number of atoms $L = 300$ (which means that the dimensions of the dictionary are 513×600), with an initial dictionary built via Online Dictionary Learning (ODL) [26] algorithm, as implemented in E-FDDL [22]. We have investigated using an initial dictionary formed by L random samples of clean speech and L random samples of noise, and we found that no gain is achieved by FDDL in this case.

All supervised speech enhancement methods need prior clean speech and noise samples training. Noise samples can be obtained either through a Voice Activity Detector (VAD) from non-speech segments, or from an offline noise database like Noisex-92 [30] (if noise recordings are available for the specific noise type we are interested in our speech enhancement system). For Noizeus, we have created a noise dataset through the subtraction of clean speech recordings from the noisy recordings in the training dataset. Regarding clean speech samples, the proposed system can use any clean speech samples we have, as we have tested the speaker independent scenario. The need for training data might look as a limitation, but we will see that it leads to a considerable superior performance compared to non-supervised speech enhancement methods that do not incorporate a training stage. To compare the performance of the proposed system with the non-supervised speech enhancement methods, we reported the performance of Geometrical Approach (GA) [2] speech enhancement method on Noizeus dataset. GA is one of the spectral subtraction method variations that addresses the problem of musical noise. We have used the implementation of GA provided by P.Loizou [31].

Table 1 shows the Frequency weighted segmental SNR for the different dictionary learning algorithms and for the GA speech enhancement method, and the improvement achieved by FDDL (FDDL gain). We can see that FDDL algorithm performs better in terms of FwSegSNR (higher values) in some cases (13 out of 27), especially in car and train noises, but not in the case of white noise, as it is not a structured noise.

Table 1:

FwSegSNR for GA method and the three DL methods, and FDDL gain

Noise	dB	GA	K-SVD	GDL	FDDL	FDDL gain%
babble	0	4.57	6.17	6.13	**6.23**	0.96
	5	5.75	7.89	**7.97**	7.91	-0.75
	10	7.40	9.76	**9.92**	9.70	-2.26
	15	9.24	11.90	**12.18**	12.13	-0.41
car	0	4.91	7.14	7.16	**7.53**	4.91
	5	6.03	8.55	8.58	**8.86**	3.16
	10	7.27	10.26	10.35	**10.63**	2.63
	15	8.79	12.57	12.69	**12.76**	0.54
restaurant	0	4.69	6.52	6.48	**6.54**	0.30
	5	6.02	7.83	**7.95**	7.86	-1.14
	10	7.61	9.66	**9.75**	9.49	-2.73
	15	9.33	11.44	**11.64**	11.46	-1.57
station	0	4.73	6.06	6.10	**6.25**	2.40
	5	5.95	7.91	8.04	**8.07**	0.37
	10	7.27	9.88	**10.02**	9.96	-0.60
	15	9.01	11.96	**12.09**	11.95	-1.17
Train	0	5.16	7.57	7.50	**7.70**	1.68
	5	6.23	8.91	8.56	**8.96**	0.55
	10	7.62	10.30	10.46	**10.74**	2.60
Airport	0	4.61	6.65	6.54	**6.74**	1.33
	5	5.75	8.16	8.28	**8.30**	0.24
	10	7.41	**10.22**	10.19	10.13	-0.88
	15	9.15	12.25	**12.31**	12.29	-0.16
White	0	4.61	**7.11**	6.98	6.92	-2.74
	5	5.71	**8.68**	8.49	8.44	-2.84
	10	6.99	**10.56**	10.26	10.28	-2.72
	15	8.36	**12.73**	12.28	12.63	-0.79

Table 2 shows the PESQ for the different dictionary learning algorithms and for the GA speech enhancement method. We notice that FDDL outperforms all other methods for all noise types regardless of the noise level, expect in the case of white noise, where K-SVD has the highest performance.

Results show that in terms of fwSegSNR, FDDL has systematically higher performance when the noise level is the highest (0 dB), while in all other cases the best approach varies. This hints that "source confusion" error has noticeable effect when noise level is high, while has a very marginal impact when noise level is low. However, in terms of PESQ, FDDL always has the highest performance for all noise types expect in the case of white noise, regardless of the noise level. This hints that "source confusion" error is more correlated with speech intelligibility than speech quality, and thus lowering this error will enhance speech intelligibility, but not necessarily speech quality. In addition, for all noise levels, FDDL has superior performance in terms of both fwSegSNR and PESQ in the case of car noise and train noise.

Table 2:

PESQ for GA method and the three DL methods, and FDDL gain

Noise	dB	GA	K-SVD	GDL	FDDL	FDDL gain%
babble	0	1.83	1.87	1.89	**1.94**	2.57
	5	2.16	2.19	2.20	**2.23**	1.34
	10	2.51	2.46	2.51	**2.52**	0.39
	15	2.83	2.76	2.85	**2.85**	0
car	0	1.84	2.24	2.28	**2.36**	3.38
	5	2.18	2.43	2.49	**2.55**	2.35
	10	2.52	2.61	2.68	**2.71**	1.10
	15	2.83	2.82	2.93	**2.93**	0
restaurant	0	1.78	1.87	1.88	**1.91**	1.57
	5	2.12	2.11	2.13	**2.17**	1.84
	10	2.48	2.44	2.47	**2.49**	0.80
	15	2.78	2.68	2.78	**2.79**	0.35
station	0	1.81	1.89	1.94	**1.98**	2.02
	5	2.18	2.23	2.29	**2.33**	1.71
	10	2.49	2.50	2.57	**2.59**	0.77
	15	2.80	2.74	2.81	**2.82**	0.35
Train	0	1.82	2.32	2.23	**2.40**	3.33
	5	2.12	2.46	2.40	**2.55**	3.52
	10	2.44	2.52	2.61	**2.74**	4.74
Airport	0	1.70	1.94	1.93	**1.99**	2.51
	5	2.17	2.25	2.26	**2.30**	1.73
	10	2.49	2.52	2.53	**2.57**	1.55
	15	2.83	2.79	2.81	**2.85**	1.40
White	0	1.79	**2.39**	2.32	2.38	-0.42
	5	2.20	**2.63**	2.54	2.61	-0.76
	10	2.53	**2.84**	2.75	2.83	-0.35
	15	2.83	**3.03**	2.95	3.03	0

We also notice that the three supervised DL based speech enhancement methods outperform the unsupervised GA speech enhancement method and have a considerable improvement over GA in terms of both fwSegSNR and PESQ.

4.3 Complexity Analysis

Vu provided a complexity analysis for E-FDDL [22], and for the original FDDL [21] denoted as O-FDDL. E-FDDL has the advantage of having considerably lower complexity compared to O-FDDL. In the context of speech enhancement, the number of classes $C = 2$ (clean speech and noise). In our experiments we have tuned the internal iteration number q for ADMM algorithm in E-FDDL dictionary update stage and found that $q = 5$ is optimal. The number of atoms in each sub-dictionary is $L = 300$, the dimension of speech feature vectors $N = 513$ and the number of training samples (frames) for both clean speech and noise $n = 1300$ (on average because we have created 10 training/test group and in each group we have different number for n), the sparsity degree $k = 30$. Rubinstein provided the implementation and complexity analysis for approximate K-SVD [20]. As we need to build two K-SVD dictionaries one for clean speech and another for noise, the complexity of using K-SVD for speech enhancement will be multiplied by 2. The same multiplication applies to the complexity of GDL, which can be calculated as follows: In each iteration we have (1) updating the dictionary using the approximate K-SVD algorithm, having a complexity of $NL^2 + 4nk(N + L) + 4NL^2$ per iteration [20], (2) updating the sparse codes using LARC algorithm (LARC is the same as LARS with different stopping criterion), having a complexity of $(N^3 + N^2 n)$ [24]. Thus, the total complexity needed for GDL is : $2(N^3 + N^2 n + NL^2 + 4nk(N + L) + 4NL^2)$. Table 3 shows the complexity analysis (per iteration) for both E-FDDL, K-SVD and GDL in the context of speech enhancement, with the plugging numbers.

Table 3:

Complexity analysis (per iteration) of using the three DL in the context of speech enhancement

Method	Complexity	Plugging numbers
K-SVD	$2(n(2NL + k^2L + 7kL + k^3 + 4kN) + 5NL^2)$	2.36×10^9
GDL	$2(N^3 + N^2 n + NL^2 + 4nk(N + L) + 4NL^2)$	1.67×10^9
E-FDDL	$C^2L((q + 1)L(N + Cn) + 2Nn)$	8.32×10^9

We notice that E-FDDL has the highest complexity, due to the internal ADMM iterations at the dictionary update stage.

5. Conclusions

In this paper we proposed a new speech enhancement system that uses FDDL for modeling both the clean speech and noise. FDDL cost function accounts for both “source distortion” and “source confusion” errors. The performance of the proposed algorithm was evaluated using two objective measures: the frequency weighted SNR (FwSegSNR) and PESQ to compare against two well-known supervised dictionary learning algorithms: K-SVD, GDL, and against GA, which is an unsupervised speech enhancement method. Experiments on Noizeus dataset shows that the unsupervised GA method has the worst performance. Results shows that FDDL has higher performance in comparison to other studied DL in terms of both measures, in almost half of the cases, but not in the case of white noise. Specifically, FDDL outperforms other studied DL methods in the presence of considerable noise (0 dB) for all studied noise types, except in the case of white noise, and in the case of structured non-stationary noise like car noise and train noise for all noise levels. Hence, it is recommended to use FDDL for speech enhancement in these cases.

Acknowledgment

The authors would like to thank P. Loizou (may God have mercy for his soul) for publishing the implementations of GA speech enhancement algorithm, and the implementation of fwSegSNR and PESQ objective measures. Also, we thank T. H. Vu for providing an excellent efficient implementation for E-FDDL.

REFERENCES

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust., Speech, Signal Process., vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] Y. Lu and P. C. Loizou, “A geometric approach to spectral subtraction,” Speech Comm., 50:453–466, 2008.

[3] J. S. Lim and V. O. Alan, “Enhancement and bandwidth compression of noisy speech,” Proceedings of the IEEE, vol. 67, no. 12, pp. 1586–1604 Dec. 1979.

[4] Y. Ephraim, “Statistical-model-based speech enhancement systems,” Proc. IEEE, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[5] Y. Hu and P. C. Loizou, “A generalized subspace approach for enhancing speech corrupted by colored noise,” IEEE Trans. Speech Audio Process., vol. 11, no. 4, pp. 334–341, Jul. 2003.

[6] J. Sun, J. Zhang, and M. Small, “Extension of the local subspace method to enhancement of speech with colored noise,” Signal Process., vol. 88, no. 7, pp. 1881–1888, Jul. 2008.

[7] D. Y. Zhao and W. B. Kleijn, “HMM-based gain modeling for enhancement of speech in noise,” IEEE Trans. Audio, Speech, and Language Process., vol. 15, no. 3, pp. 882–892, Mar. 2007.

[8] N. Mohammadiha, R. Martin, and A. Leijon, “Spectral domain speech enhancement using HMM state-dependent super-Gaussian priors,” IEEE Signal Process. Letters, vol. 20, no. 3, pp. 253–256, Mar. 2013.

[9] H. Veisi and H. Sameti, “Speech enhancement using hidden Markov models in Mel-frequency domain,” Speech Communication, vol. 55, no. 2, pp. 205–220, Feb. 2013.

[10] K. W. Wilson, B. Raj, and P. Smaragdis, “Regularized non-negative matrix factorization with temporal dependencies for speech denoising,” in Proc. Int. Conf. Spoken Language Process. (Interspeech), 2008, pp. 411–414.

[11] M. Sun, Y. Li, J. Gemmeke, and X. Zhang, “Speech enhancement under low snr conditions via noise estimation using sparse and low-rank nmf with kullback-leibler divergence,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 7, pp. 1233–1242, Jul. 2015.

[12] N. Mohammadiha, P. Smaragdis, and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” IEEE Trans. Audio, Speech, Lang. Process., vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[13] C. D. Sigg, T. Dikk, and J. M. Buhmann, “Speech enhancement using generative dictionary learning,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 6, pp. 1698–1712, Aug. 2012.

[14] Y. Zhao, X. Zhao, B. Wang, “A speech enhancement method based on sparse reconstruction of power spectral density,” Computers & Electrical Engineering, Volume 40, Issue 4, 2014, Pages 1080-1089, ISSN 0045-7906.

[15] Y. Luo, G. Bao, Y. Xu and Z. Ye, “Supervised Monaural Speech Enhancement Using Complementary Joint Sparse Representations,” IEEE Signal Processing Letters, vol. 23, no. 2, pp. 237-241, Feb. 2016.

[16] L. Zhang, G. Bao, J. Zhang, Z. Ye, “Supervised single-channel speech enhancement using ratio mask with joint dictionary learning,” Speech Communication, vol. 82, pp. 38-52, Sep. 2016.

[17] K. Engan, S. O. Aase, and J. Hakon Husoy, “Method of optimal directions for frame design,” Proceedings of IEEE ICASSP, 1999, vol. 5

[18] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” IEEE Trans. Image Process., vol. 15, no. 12, pp. 3736-3745, Dec. 2006.

[19] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” IEEE Trans. on Signal Processing, vol. 54, no. 11, pp. 4311–4322, 2006.

[20] R. Rubinstein, M. Zibulevsky, and M. Elad. Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit. Technical report, Technion, Haifa, 2008.

[21] M. Yang, L. Zhang, X. Feng, “Sparse Representation Based Fisher Discrimination Dictionary Learning for Image Classification,” International Journal of Computer Vision, 2014.

[22] T. H. Vu and V. Monga, "Fast Low-Rank Shared Dictionary Learning for Image Classification," in IEEE Transactions on Image Processing, vol. 26, no. 11, pp. 5160-5175, Nov. 2017.

[23] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition,” Proc. Asilomar Conf. Signal Syst. Comput., 1993.

[24] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” Ann. Stat., 32:407–499, 2004.

[25] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” SIAM Journal on Imaging Sciences, vol. 2, no. 1, pp. 183–202, 2009.

[26] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online learning for matrix factorization and sparse coding,” The Journal of Machine Learning Research, vol. 11, pp. 19–60, 2010.

[27] NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms:

http://ecs.utdallas.edu/loizou/speech/noizeus/

[28] 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,” Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process., pages 749–752, 2001.

[29] P. C. Loizou, Speech Enhancement: Theory and Practice. Boca Raton, FL, USA: CRC, 2013.

[30] Noisex-92: Database of recording of various noises:

www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

[31] http://ecs.utdallas.edu/loizou/speech/software.htm

[32] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Foundations and Trends ® in Machine Learning, vol. 3, no. 1, pp. 1– 122, 2011.

[33] M. Yang, L. Zhang, J. Yang, D. Zhang, “Metaface learning for sparse representation-based face recognition,” Proc. IEEE Conference on Image Processing, 2010.

[34] M. Yang, L. Zhang, X. Feng, D. Zhang, “Fisher discrimination dictionary learning for sparse representation,” Proceedings of the International Conference on Computer Vision, 2011.

[35] S. Zhang, X. Zhao, Y. Chuang, W. Guo and Y. Chen, “Learning Discriminative Dictionary for Facial Expression Recognition,” IETE Technical Review, 2017. DOI: 10.1080/02564602.2017.1283251