MULTICHANNEL TRANSIENT ACOUSTIC SIGNAL CLASSIFICATION USING TASK-DRIVEN DICTIONARY WITH JOINT SPARSITY AND BEAMFORMING

Yang Zhang[†] *Nasser M. Nasrabadi*[⋆] *Mark Hasegawa-Johnson*[†]

[†]University of Illinois, Urbana-Champaign, Department of Electrical and Computer Engineering
[⋆]U.S. Army Research Laboratory, Adelphi

yzhan143@illinois.edu nasser.m.nasrabadi.civ@mail.mil jhasegaw@illinois.edu

ABSTRACT

We are interested in a multichannel transient acoustic signal classification task which suffers from additive/convolutionary noise corruption. To address this problem, we propose a double-scheme classifier that takes the advantage of multichannel data to improve noise robustness. Both schemes adopt task-driven dictionary learning as the basic framework, and exploit multichannel data at different levels - scheme 1 imposes joint sparsity constraint while learning the dictionary and classifier; scheme 2 adopts beamforming at signal formation level. In addition, matched filter and robust ceptral coefficients are applied to improve noise robustness of the input feature. Experiments show that the proposed classifier significantly outperforms the baseline algorithms.

Index Terms— Transient acoustic signal, multichannel, task-driven dictionary learning, joint sparsity, beamforming

1. INTRODUCTION

Transient acoustic signal classification is an important task in battlefield surveillance [1–3]. Specifically, we are interested in a two-class classification of artillery explosion. This task comes with several major challenges. First, the classification is usually performed under noisy environment, and the noise level varies drastically. Second, the acoustic event usually takes place in an unknown environment without realtime calibration of channel characteristic. Third, transient acoustic signal only lasts for a very short period of time. Therefore, each acoustic event has only a few samples, typically no more than 1000, available for feature extraction. In addition, unlike speech or music signals, which exhibit short-time stationarity, transient signals are unsuitable for framed analysis.

Dictionary learning with sparse codes is a popular approach for denoising [4], noise robust classification [5–8], and other tasks [9]. A signal usually exhibits sparse pattern in some transformed space. By properly transforming the noisy signal and projecting it onto the $\mathcal{L}_0$ ball (remove small dimensions), one can effectively remove the noise energy. While many classification schemes based on sparse codes are generative [5–8], i.e., a dictionary is trained for each class, and a sample is classified to a class whose dictionary can reconstruct it most accurately, Mairal et.al. [10] proposed a task-driven dictionary learning scheme that learns the dictionary discriminatively. Given its noise robustness and discriminative natures, task driven dictionary is an ideal framework for our task.

This research was supported in part by ARO grant W911NF-09-1-0383 and AHRQ grant R21HS022948. All results and opinions are those of the authors and are not endorsed by ARO.

The presence of multiple channel (sensors) provides further flexibility for our classification task. In our task, the acoustic events were recorded with a tetrahedral sensor whose configuration is known. With the tetrahedral sensor, we can now leverage the dependency among different channels to improve noise robustness and/or augment feature dimension. The fusion of multichannel data can be performed at the signal formation level or learning feature level.

At the signal formation level, beamforming is a traditional and effective algorithm to obtain a more accurate estimate of the signal for various applications [11,12]. Since the transient acoustic signal is wideband, we will apply an enhanced version of MVDR [13], which divides wideband signal into narrow bands and performs beamforming within each band. The enhanced version takes into account the channel differences and estimates the convolutionary noise.

At the learning feature level, joint sparsity constraint is widely used in improving noise robustness of the sparse code [14–16]. Joint sparsity enforces the sparse codes of different channels of the same acoustic event to have the same sparse pattern, i.e., same non-zero dimensions. It improves noise robustness because it projects noise onto a more constrained set while the clean signal is unaffected. In addition, the dependency between the outputs of the channel classifiers can be further exploited by introducing a regularization term that penalizes inter-channel differences in classification outputs.

This paper proposes a classifier that is carefully tailored to this specific task. The classifier stacks two learning schemes. The first scheme uses the task driven dictionary learning with joint sparsity constraint and a regularization on the channel classifier outputs. The second scheme estimates the noise-robust feature using enhanced MVDR and feeds it into the task-driven dictionary. In addition, the proposed method incorporates other important techniques such as MMSE cepstral estimation [17] and robust acoustic event detection using matched filter [18]. Experiments show that the proposed algorithm outperforms traditional approaches significantly, especially when dataset is large with various noise levels.

There have been previous efforts to solve the multichannel acoustic signal classification. Zhang et. al. [19] proposed an algorithm that builds a structured dictionary with joint sparsity for each class, and assigns the test sample to the class whose dictionary has least reconstruction error. Srinivas et. al. [20] designed a graph-based method that learns probability dependence among different channels from the data. However, both algorithms were designed for very small datasets, and noise variation was limited. With a much larger dataset now available, noise variability and test complexity become major challenges. The proposed method in this paper addresses these challenges.

The remainder of the paper is organized as follows. Section 2 gives an overview of the structure of the proposed classifier. Section

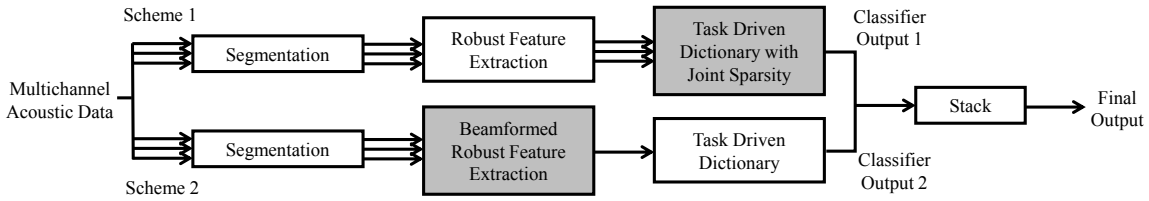

Fig. 1: Overall structure of the classifier.
Triple arrow indicates multichannel data/feature; single arrow indicates single channel data/feature. Grey box indicates where multichannel fusion of takes place.

3 briefly introduces task driven dictionary learning and adapts the training formula for the joint sparsity case. Section 4 introduces the enhanced version of MVDR and MMSE cepstral coefficient estimation. It also discusses the theoretical justification for combining both algorithms. Section 5 evaluates the performance of the classifier. Section 6 concludes the paper and points out some future directions.

2. CLASSIFIER STRUCTURE OVERVIEW

Fig. 1 shows the overall structure of the classifier. As can be seen, the classifier consists of two learning schemes. It linearly combines the two outputs to produce the final classification result. Scheme 1 extracts the event segments using a matched filter (box labeled "segmentation"), and estimates noise-robust cepstral coefficients from each channel using MMSE estimation (box labeled "Robust Extraction"). The multichannel features are then fed into the task driven dictionary learning with joint sparsity constraint for classification.

In scheme 2, the multi-channel data are fused into one enhanced signal, from which robust cepstral coefficients are estimated using MMSE estimation. This uni-channel feature is then fed into the classical task-driven dictionary learning scheme for classification.

Both schemes apply task-driven dictionary learning as framework, and both fuse the multichannel information to further improve noise robustness. The difference is that in scheme 1, the fusion is done at dictionary learning stage, while in scheme 2 it is done in the signal formation level. In the following two sections we will focus on dictionary learning and feature extraction respectively. Particularly, we will discussion how the fusion is done at these two stages.

3. TASK-DRIVEN DICTIONARY LEARNING

3.1. Classical Task-Driven Dictionary Learning Scheme

We will use the task-driven dictionary learning proposed in [10] as the framework for scheme 2. Let $\boldsymbol{x}$ denote the input feature, which, in this case, is a column vector of dimension M, and let y denote its class (0 or 1). Let M-by-N matrix $\boldsymbol{D}$ denote the dictionary. The N dimensional sparse code $\boldsymbol{\alpha}$, as a function of $\boldsymbol{x}$ and $\boldsymbol{D}$, is obtained by

$$\boldsymbol{\alpha}^*(\boldsymbol{x},\boldsymbol{D}) = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{\alpha}\|_2^2, \quad (1)$$

where λ_1 and λ_2 are regularization parameters.

The sparse code $\boldsymbol{\alpha}^*(\boldsymbol{x},\boldsymbol{D})$ is then fed into a linear classifier defined by $\boldsymbol{w}$. $\boldsymbol{w}$ and $\boldsymbol{D}$ are determined by minimizing the expected loss with an $\mathcal{L}_2$ regularization

$$\boldsymbol{w},\boldsymbol{D} = \underset{\boldsymbol{w}',\boldsymbol{D}'}{\operatorname{argmin}} f(\boldsymbol{w}',\boldsymbol{D}') + \frac{\nu}{2}\left\|\boldsymbol{w}'\right\|_2^2, \quad (2)$$

where ν is a regularization parameter and $f(\boldsymbol{w},\boldsymbol{D})$ is the expected loss function

$$f(\boldsymbol{w},\boldsymbol{D}) = E_{y,\boldsymbol{x}}\left[l\left(y,\boldsymbol{w}^T\boldsymbol{\alpha}^*(\boldsymbol{x},\boldsymbol{D})\right)\right], \quad (3)$$

and $l(y,\hat{y})$ is some differentiable loss function of the linear classifier.

It is proved in [10] that $f(\boldsymbol{w},\boldsymbol{D})$ is differentiable with respect to $\boldsymbol{w}$ and $\boldsymbol{D}$, and it is easy to derive

$$\begin{aligned}\nabla_{\boldsymbol{w}} f &= E_{y,\boldsymbol{x}}\left[\nabla_{\boldsymbol{w}} l(y,\boldsymbol{w}^T\boldsymbol{\alpha}^*(\boldsymbol{x},\boldsymbol{D}))\right]\\ \nabla_{\boldsymbol{D}} f &= E_{y,\boldsymbol{x}}\left[\boldsymbol{D}\boldsymbol{\beta}^*\boldsymbol{\alpha}^{*T} + (\boldsymbol{x}-\boldsymbol{D}\boldsymbol{\alpha}^*)\boldsymbol{\beta}^{*T}\right],\end{aligned} \quad (4)$$

where

$$\begin{aligned}\boldsymbol{\beta}_{\boldsymbol{\Lambda}^C} &= 0\\ \boldsymbol{\beta}_{\boldsymbol{\Lambda}} &= \left(\boldsymbol{D}_{\boldsymbol{\Lambda}}^T\boldsymbol{D}_{\boldsymbol{\Lambda}} + \lambda_2\boldsymbol{I}\right)^{-1}\left[\nabla_{\boldsymbol{\alpha}_{\boldsymbol{\Lambda}}} l(y,\boldsymbol{w}^T\boldsymbol{\alpha}^*(\boldsymbol{x},\boldsymbol{D}))\right],\end{aligned} \quad (5)$$

and $\boldsymbol{\Lambda}$ is a set of indices of nonzero elements of $\boldsymbol{\alpha}$. By careful initialization and using the standard gradient descent method, we can achieve a satisfactory local optimum.

3.2. Task-Driven Dictionary Learning with Multiple Channels

In the presence of multichannel data, the observation measurements become an M-by-K matrix, $\boldsymbol{X}$, where each column represents a channel and K is the total number of channels. Accordingly, the sparse codes $\boldsymbol{\alpha}$ become an N-by-K matrix. Compared to the classical task-driven dictionary, the proposed multichannel task-driven dictionary comes with two major adaptations.

First, the sparsity regularization term on $\boldsymbol{\alpha}$ becomes a joint sparsity regularization

$$\|\boldsymbol{\alpha}\|_{1\backslash 2} = \sum_{k=1}^{K}\|\boldsymbol{\alpha}_{k,:}\|_2, \quad (6)$$

where $\boldsymbol{\alpha}_{k,:}$ is the k-th row of $\boldsymbol{\alpha}$. Eq. (6) would force the sparse codes of all the channels to have same non-zero components, which is reasonable because in the ideal noise-free case, if the feature is selected appropriately, the features of the same event extracted from different channels should be exactly the same.

Second, an additional regularization is imposed on the channel classification results. With K different channels, the classifier output, $\boldsymbol{w}^T\boldsymbol{\alpha}$, now becomes a 1-by-$K$ row vector. The final classification decision is made based on the mean output of the K channels, mean($\boldsymbol{w}^T\boldsymbol{\alpha}$). In order to reduce variation of classification outputs induced by noise, another regularization term is added to the target function so that the outputs of all the channel classifiers in

$\boldsymbol{w}^T\boldsymbol{\alpha}^*(\boldsymbol{X},\boldsymbol{D})$ are similar.

Formally, the proposed minimization problem now has two regularization terms, as shown by the second and third terms in Eq. (7)

$$\boldsymbol{w},\boldsymbol{D} = \underset{\boldsymbol{w}',\boldsymbol{D}'}{\operatorname{argmin}} f(\boldsymbol{w}',\boldsymbol{D}') + \frac{\nu_1}{2}\left\|\boldsymbol{w}'\right\|_2^2 + \frac{\nu_2}{2} g\left(\boldsymbol{w}'^T\boldsymbol{\alpha}^*(\boldsymbol{X},\boldsymbol{D}')\right), \tag{7}$$

where the expected loss function is still similar to (3):

$$f(\boldsymbol{w},\boldsymbol{D}) = E_{y,\boldsymbol{X}}\left[l\left(y, \text{mean}\left(\boldsymbol{w}^T\boldsymbol{\alpha}^*(\boldsymbol{X},\boldsymbol{D})\right)\right)\right], \tag{8}$$

and $g(\cdot)$ calculates the expected summation of pairwise squared Euclidean distance between the elements of the vector.

Similar to (1), the sparse code is calculated by

$$\boldsymbol{\alpha}^*(\boldsymbol{x},\boldsymbol{D}) = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{x}-\boldsymbol{D}\boldsymbol{\alpha}\right\|_2^2 + \lambda_1\left\|\boldsymbol{\alpha}\right\|_{1\backslash 2} + \frac{\lambda_2}{2}\left\|\boldsymbol{\alpha}\right\|_F^2, \tag{9}$$

where $\|\cdot\|_F$ is the Frobenius norm. Readers can contrast Eqs. (1)-(3) and (7)-(9) to apprehend the differences between the two schemes.

With the new scheme, the gradients become much more complicated. The $\boldsymbol{\beta}$ in (5) becomes an N-by-K matrix:

$$\begin{aligned}
&\boldsymbol{\beta}_{\boldsymbol{\Lambda}^C,:} = 0\\
&\text{vec}\left(\boldsymbol{\beta}_{\boldsymbol{\Lambda},:}\right) = \left(\text{kron}\left(\boldsymbol{I}_K, \boldsymbol{D}_{\boldsymbol{\Lambda}}^T\boldsymbol{D}_{\boldsymbol{\Lambda}}\right) + \lambda_2\boldsymbol{I}_{MK} + \lambda_1\boldsymbol{A}\right)^{-1}\cdot\\
&\left[\nabla_{\text{vec}(\boldsymbol{\alpha}_{\boldsymbol{\Lambda},:})} l\left(y, \text{mean}\left(\boldsymbol{w}^T\boldsymbol{\alpha}^*(\boldsymbol{X},\boldsymbol{D})\right)\right)\right],
\end{aligned} \tag{10}$$

where subscript $\boldsymbol{\Lambda},:$ denotes the submatrix consisting of rows specified by the set $\boldsymbol{\Lambda}$; I_K denotes K-by-K identity matrix; vec$(\cdot)$ denotes vectorization, i.e., turning the matrix into a column vector by concatenating its columns; kron$(\cdot,\cdot)$ denotes Kronecker product.

The super matrix $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{B}^3\boldsymbol{C}, \tag{11}$$

where

$$\boldsymbol{B} = \text{kron}\left[\boldsymbol{I}_K, \underset{i}{\text{diag}}\left(\left\|\boldsymbol{\alpha}^*_{:,i}(\boldsymbol{X},\boldsymbol{D})\right\|_2^{-1}\right)\right], \tag{12}$$

$\underset{i}{\text{diag}}(\cdot)$ is a diagonal matrix whose entry (i,i) is given by the argument; $\boldsymbol{C}$ is a block matrix with K row partitions and K column partitions. The (i,j)th block is

$$\boldsymbol{C}_{ij} = \underset{k}{\text{diag}}\left(\alpha^*_{k,i}(\boldsymbol{X},\boldsymbol{D})\alpha^*_{k,j}(\boldsymbol{X},\boldsymbol{D})\right). \tag{13}$$

Inverting a super large matrix increases the computational complexity of dictionary learning with the joint sparsity constraint by K^3 with respect to the classical version. Nevertheless, we can still obtain a local optimal solution by the gradient descent method.

4. NOISE ROBUST FEATURE EXTRACTION

4.1. Acoustic Event Detection

Before feature extraction, a segment that contains the acoustic signal of the event is cut out from the original recording. In previous work [19, 20], this is done by spectral maximum detection [21], which essentially passes the signal through a 1st-order difference filter.

Having more knowledge about the signal, we can further improve the SNR by using a matched filter [18]. Previous studies of the artillery explosion signal [22] have shown that, despite the drastic differences in 'whiskers' and explosive burning, different explosions have a similar shape of main blast. We manually label a small number of main blast signals randomly drawn from the training set, and average over them. The resulting signal is then applied as the impulse response of the matched filter.

4.2. Noise-Robust Cepstral Coefficients

Cepstral coefficients have been proven effective in speech and acoustic signal processing [23] tasks. One of the major drawbacks of cepstral coefficients is that additive noise corrupts all of the cepstral coefficients [24]. To improve noise robustness, we apply the MMSE cepstral estimate proposed in [17].

Denote the clean signal spectrum by $\boldsymbol{s}$; then the vector of cepstral coefficients $\boldsymbol{x}$ is given by

$$\boldsymbol{x} = \text{DCT}\left(\log|\boldsymbol{s}|\right), \tag{14}$$

where DCT$(\cdot)$ denotes discrete cosine transform operation.

Suppose the observation is corrupted by additive noise, $\boldsymbol{n}$. We know that MMSE estimate of $\boldsymbol{x}$ is its posterior expectation:

$$\hat{\boldsymbol{x}} = E[\boldsymbol{x}|\boldsymbol{s}+\boldsymbol{n}] = \text{DCT}\left(E\left[\log|\boldsymbol{s}|\,|\,\boldsymbol{s}+\boldsymbol{n}\right]\right). \tag{15}$$

The second equality holds because DCT is a linear operation.

In [17] authors derived that if both $\boldsymbol{s}$ and $\boldsymbol{n}$ are jointly independent Gaussian, with the i-th dimension being $\mathcal{N}\left(\mu_i,\sigma_i^2\right)$ and $\mathcal{N}\left(0,\lambda_i^2\right)$ respectively, then

$$\begin{aligned}
&E\left[\log|s_i|\,|\,\boldsymbol{s}+\boldsymbol{n}\right] =\\
&\log\left[\frac{\xi_i}{1+\xi_i}\right] + \frac{1}{2}\int_{\nu_i}^{\infty}\frac{e^{-t}}{t}dt + \log\left|s_i+n_i\right|,
\end{aligned} \tag{16}$$

where s_i and n_i are the i-th element of $\boldsymbol{s}$ and $\boldsymbol{n}$, and

$$\nu_i = \frac{\xi_i}{1+\xi_i}\gamma_i; \quad \xi_i = \frac{\sigma_i^2}{\lambda_i^2}; \quad \gamma_i = \frac{|s_i+n_i|^2}{\lambda_i^2}, \tag{17}$$

ξ_i and γ_i can be interpreted as prior and posterior SNRs respectively.

We assume the prior variance of the signal, σ_i^2, to be uniform across all i, and is equal to the mean signal energy per frequency bin. This is estimated from the training data. As shown in figure 1, scheme 1 extracts noise-robust cepstral coefficients for each channel, keeps the low quefrencies, and sends the multichannel feature to the task-driven dictionary, where multichannel data are fused.

4.3. Beamformed Cepstral Coefficients

In scheme 2, as shown in Fig. 1, fusion is done at signal formation level, where only 1 set of cepstral coefficients is extracted out from multichannel data. In this case, each channel k adds a different noise, $\boldsymbol{n}^{(k)}$, to the same clean signal, $\boldsymbol{s}$, and the estimate becomes

$$\hat{\boldsymbol{x}} = E[\boldsymbol{x}|\{\boldsymbol{s}+\boldsymbol{n}^{(k)}\}_k] = \text{DCT}\left(E\left[\log|\boldsymbol{s}||\{\boldsymbol{s}+\boldsymbol{n}^{(k)}\}_k\right]\right), \tag{18}$$

where $\{\boldsymbol{s}+\boldsymbol{n}^{(k)}\}_k$ denotes the set of multichannel noisy observations. Similar to the derivation in [25], it can be shown that solving (18) breaks into two steps.

By definition, the posterior probability of $\boldsymbol{s}$ can be re-expressed in terms of sufficient statistics

$$p\left(\boldsymbol{s}|\{\boldsymbol{s}+\boldsymbol{n}^{(k)}\}\right) = p\left(\boldsymbol{s}|T\left(\{\boldsymbol{s}+\boldsymbol{n}^{(k)}\}_k\right)\right), \tag{19}$$

where $T\left(\{\boldsymbol{s}+\boldsymbol{n}^{(k)}\}_k\right)$ are the sufficient statistics for $\boldsymbol{s}$. It is straightforward that a similar equality holds for $\log|\boldsymbol{s}|$, and therefore

$$E\left[\log|\boldsymbol{s}||\{\boldsymbol{s}+\boldsymbol{n}^{(k)}\}_k\right] = E\left[\log|\boldsymbol{s}||T\left(\{\boldsymbol{s}+\boldsymbol{n}^{(k)}\}_k\right)\right]. \quad (20)$$

Moreover, it has been shown in [25] that if we assume the noises of all channels are jointly Gaussian then $T\left(\{\boldsymbol{s}+\boldsymbol{n}^{(k)}\}_k\right)$ is the beamformed signal estimated by MVDR [11], which is still Gaussian. Therefore, the right hand side of (20) can be calculated the same way as in (16).

In short, estimating cepstral coefficients $\hat{\boldsymbol{x}}$ involves two steps. First, obtain the beamformed signal from the multichannel signal by MVDR. Second, obtain noise-robust cepstral coefficients of the beamformed signal as in section 4.2. To correct for convolutionary noise, and account for different SNR's and gains of each channel, we apply an enhanced MVDR algorithm proposed in [13].

5. EXPERIMENT RESULTS

5.1. Dataset and Configurations

The dataset we experimented with is collected by the US Army Research Lab. It contains 3941 four-channel event samples of explosion sound of a certain type of artillery recorded by the tetrahedral array. Our task is to classify if the explosion is a launch or an impact. Data were collected at 5 different sites, which have different types of additive and convolutionary noises. The range of SNR of the data set is very wide — from -10 dB to 50 dB. The presence of site-dependent convolutionary noise makes the problem even harder.

We extract 50-dimensional cepstral features and limit the number of atoms in the dictionary to 30. Squared loss is applied as the loss function l. For each of the following experiment, we apply 5-fold cross validation with training ratio being 0.5. To compute the sparse code as in Eqs. (1) and (9), we apply the FISTA [26] algorithm with 300 iterations; to compute the optimal classifier and dictionary using Eqs. (2) and (7), we apply the simple gradient descent approach with 1500 iterations.

5.2. General Results

For comparison purposes, we also look at the performance of SVM with RBF kernel as our baseline method. We use SVM clssifier with two different input scenarios. In the first scenario, denoted as SVM beamform, we use beamformed noise-robust cepstral coefficients as input; The second, denoted as SVM concat, uses concatenated noise-robust cepstral coefficients of the 4 channels as input. To see how much each part of the algorithm contributes to the overall performance, we also compare several variants of our proposed algorithms, where a part of the algorithm is removed.

Table 1 displays the results, where the uppermost panel is the general result. As can be seen, both schemes are significantly better than the baseline, and linearly combining the two schemes further improves the performance. Between the two schemes, scheme 1 works better. Notice that scheme 1 has 3 times more training samples than scheme 2, and features of scheme 2 are cleaner. This results indicates that data size is a more binding limit.

5.3. Performance Decomposition

Now, we will discuss how each part of the proposed algorithm would contribute to the overall performance. The third sub-panel in table 1 lists the results where joint sparsity constraint or beamforming is removed. The mix channel is a variant of scheme 1 where the sparse code is no longer constrained to have the same sparse pattern while the rest are the same as scheme 1. Single channel is a variant of scheme 2 where only channel 1 signal is fed into cepstral coefficients estimation instead of the beamformed signal. As can be seen, mix channel works quite satisfactorily. Although the training accuracy is significantly lower than scheme 1, the deterioration does not generalize as much to test set. This is because mix channel scheme still fuses multichannel to a very high degree by having different channels share the same dictionary and regularizing the classification output of different channels to be similar. With larger training set, however, we expect the improvement brought by joint sparsity would become more significant. On the other hand, single channel works much poorer than scheme 2, which is reasonable because the noise robustness provided by multichannel fusion is completely removed.

Table 1: Performance of proposed algorithm and its variants

Algorithm	Training Accuracy	Test Accuracy
scheme 1	95.76	83.63
scheme 2	92.46	82.07
final	96.25	84.39
Kernel SVM baselines		
SVM beamform	99.64	77.38
SVM concat	99.47	77.20
Algorithms without joint sparsity or beamforming		
mix channel	93.18	83.46
single channel	91.38	80.66
Algorithms with normal cepstral coefficients		
scheme 1	95.90	81.58
scheme 2	90.99	78.62
Algorithms without task-driven dictionary		
scheme 1	75.88	74.99
scheme 2	76.81	75.57

The fourth panel shows the performance of variants where normal cepstral coefficients instead of the noise-robust ones are applied as input features. As can be seen, both schemes degrade significantly. Particularly, the gap between training and test accuracies are much greater, indicating noise-robust cepstral coefficients help in reducing noise.

The last panel displays the performance when the dictionary is trained with the minimum reconstruction error criterion instead of task-driven dictionary learning criterion. We see there is a large performance degradation in both schemes, as was explained in [10]. The reason for this degradation is that the dictionary is no longer discriminative.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose a double-scheme algorithm which combines task-driven dictionary with joint sparsity and beamforming, and is further strengthened by many noise robust algorithms. Experiment shows that the proposed algorithm improves over baseline algorithms. A drawback, though, is that the classifier we apply is a simple linear classifier, while many research efforts verify that nonlinear classifier would further improve the performance. How to incorporate nonlinear classifier into our framework would be our future direction.

7. REFERENCES

[1] MR Azimi-Sadjadi, Y Jiang, and S Srinivasan, "Acoustic classification of battlefield transient events using wavelet sub-band features," in *Defense and Security Symposium*. International Society for Optics and Photonics, 2007, pp. 656215–656215.

[2] B Kaushik, Don Nance, and KK Ahuja, "A review of the role of acoustic sensors in the modern battlefield," 2005.

[3] Vincent Mirelli, Stephen Tenney, Yoshua Bengio, Nicolas Chapados, and Olivier Delalleau, "Statistical machine learning algorithms for target classification from acoustic signature," *Proceedings of MSS Battlespace Acoustic and Magnetic Sensors*, 2009.

[4] B. Wen, S. Ravishankar, and Y. Bresler, "Structured overcomplete sparsifying transform learning with convergence guarantees and applications," *International Journal of Computer Vision.*, 2014, Submitted for publication.

[5] Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3501–3508.

[6] Meng Yang, D Zhang, and Xiangchu Feng, "Fisher discrimination dictionary learning for sparse representation," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 543–550.

[7] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[8] Pablo Sprechmann and Guillermo Sapiro, "Dictionary learning and sparse coding for unsupervised clustering," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2042–2045.

[9] Shiyu Chang, Charu C. Aggarwal, and Thomas S. Huang, "Learning local semantic distances with limited supervision," in *Data Mining (ICDM), 2014 IEEE International Conference on*, Dec 2014, pp. 70–79.

[10] Julien Mairal, Francis Bach, and Jean Ponce, "Task-driven dictionary learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 791–804, 2012.

[11] Lloyd J Griffiths and Charles W Jim, "An alternative approach to linearly constrained adaptive beamforming," *Antennas and Propagation, IEEE Transactions on*, vol. 30, no. 1, pp. 27–34, 1982.

[12] Osamu Hoshuyama, Akihiko Sugiyama, and Akihiro Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *Signal Processing, IEEE Transactions on*, vol. 47, no. 10, pp. 2677–2684, 1999.

[13] Demba E Ba, Dinei Florêncio, and Cha Zhang, "Enhanced MVDR beamforming for arrays of directional microphones," in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 1307–1310.

[14] Ewout van den Berg and Michael P Friedlander, "Joint-sparse recovery from multiple measurements," *arXiv preprint arXiv:0904.2051*, 2009.

[15] Marco F Duarte, Shriram Sarvotham, Michael B Wakin, Dror Baron, and Richard G Baraniuk, "Joint sparsity models for distributed compressed sensing," in *Proceedings of the Workshop on Signal Processing with Adaptative Sparse Structured Representations*, 2005.

[16] Moshe Mishali and Yonina C Eldar, "Reduce and boost: Recovering arbitrary sets of jointly sparse vectors," *Signal Processing, IEEE Transactions on*, vol. 56, no. 10, pp. 4692–4702, 2008.

[17] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

[18] D Oo North, "An analysis of the factors which determine signal/noise discrimination in pulsed-carrier systems," *Proceedings of the IEEE*, vol. 51, no. 7, pp. 1016–1027, 1963.

[19] Haichao Zhang, Yanning Zhang, Nasser M Nasrabadi, and Thomas S Huang, "Joint-structured-sparsity-based classification for multiple-measurement transient acoustic signals," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 6, pp. 1586–1598, 2012.

[20] Umamahesh Srinivas, Nasser M Nasrabadi, and Vishal Monga, "Graph-based sensor fusion for classification of transient acoustic signals," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 2014.

[21] Shlomo Engelberg and Eitan Tadmor, "Recovery of edges from spectral data with noise-a new perspective," *SIAM Journal on Numerical Analysis*, vol. 46, no. 5, pp. 2620–2635, 2008.

[22] Myron E Hohil, Sachi Desai, and Amir Morcos, "Reliable classification of high explosive and chemical/biological artillery using acoustic sensors," Tech. Rep., DTIC Document, 2006.

[23] Thomas F Quatieri, *Discrete-time speech signal processing: principles and practice*, Pearson Education India, 2002.

[24] Arthur Nádas, David Nahamoo, and Michael A. Picheny, "Speech recognition using noise-adaptive prototypes," vol. 37, no. 10, pp. 1495–1503, 1989.

[25] Radu Balan and Justinian Rosca, "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*. IEEE, 2002, pp. 209–213.

[26] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.