Cosine similarity based dictionary learning and source recovery for classification of diverse audio sources

K V Vijay Girish
Department of Electrical Engineering,
Indian Institute of Science, Bangalore
Email: kv@ee.iisc.ernet.in

T V Ananthapadmanabha
Voice and Speech Systems,
Malleswaram, Bangalore, India
Email: tva.blr@gmail.com

A G Ramakrishnan
Department of Electrical Engineering,
Indian Institute of Science, Bangalore
Email: ramkiag@ee.iisc.ernet.in

***Abstract*—A dictionary learning based audio source classification algorithm is proposed. Cosine similarity measure is used to select the atoms during dictionary learning. Three proposed objective measures, namely, signal to distortion ratio (SDR), the number of non-zero weights and the sum of weights have been used for classification. A frame-wise source classification accuracy of 98.86% is obtained for twelve different sources using SDR measure and a secondary support vector machine classifier. 100% accuracy has been obtained using moving SDR accumulated over 14 successive frames. For ten of the audio sources tested, 100% accuracy requires accumulation of only 6 frames of a signal.**

Index Terms: Dictionary learning, cosine similarity, audio classification, source recovery, sparse representation.

I. Introduction

A. Motivation for the present study

The nature of noise in an audio signal varies with the environment such as traffic, restaurant, railway and bus station. Even competing speakers and music may impair intelligibility of speech. In the case of speech enhancement [1] and noise source separation, especially for hearing impaired [2], [3], the suppression of background audio for improving the intelligibility of speech would be more effective, if the type of background audio source can be classified. Other interesting applications of noise identification are forensics [4], machinery noise diagnostics [5], robotic navigation systems [6] and acoustic signature classification of aircrafts or vehicles [7].

Noise classification can be seen as a first step in machine listening [8], which enables the system to know the background environment. Classification of noise types has been reported in the case of pure noise sources. Kates [9] addressed the problem of noise classification for hearing aid applications based on the variation of signal envelope as features. Maleh et al. [10] used line spectral frequencies as features for classification of different kinds of noise as well as noise and speech classification. Chu et al. [12] recognized fourteen different environmental sounds using matching pursuit based features combined with mel-frequency cepstral coefficients. Liu et. al. [13] devised a TV broadcast video classifier using hidden Markov model (HMM) with audio features. Zhang et. al. [14] and Lu et. al. [15] segmented and classified audio signals using statistical analysis of simple audio features and a rule-based classifier. Ma et. al. [16], [17] and Couvreur et. al. [18] devised a HMM based noise classifier for context awareness. Cherla et. al. [19] and Ramasubramanian et. al.[20] proposed a novel technique for audio analytics and audio indexing using template based modeling of audio classes and HMMs. Ramasubramanian et. al. [21] addressed the problem of audio indexing into a target and background class using Gaussian mixture models. Giannoulis et al. [22] conducted a public evaluation challenge on acoustic scene classification (similar to noise classification), where eleven algorithms were evaluated along with a baseline system. The algorithms use time and frequency domain features extracted from the audio signal followed by a statistical model or majority vote based classifier. Cauchi [23] used non-negative matrix factorization for classification of auditory scenes.

This paper addresses the basic problem of classification of the type of audio from a finite set of sources. Representation of audio signals as a sparse, linear combination of non-negative vectors called as dictionary atoms has been used for audio source separation [24], [25], [26], recognition [27], [28], classification [29], [30] and coding [31], [32]. In this work, we only address the problem of audio classification of clean noise sources using sparse non-negative representation of audio by proposing a novel dictionary learning and a source recovery method. However, the proposed audio classification also works with a mixed audio signal, where segments have higher noise energy than speech.

B. Review of dictionary learning and source recovery

A dictionary is a matrix $\mathbf{D} \in \mathbb{R}^{\mathrm{P}\times\mathrm{K}}$ (with P as the dimension of the acoustic feature vector) containing K column vectors called atoms, denoted as $\mathbf{d}_n, 1 \leq n \leq K$. Any real valued feature vector can be represented as $\mathbf{y} \approx \mathbf{Dx}$, where $\mathbf{x} \in \mathbb{R}^{\mathrm{K}}$ is the vector containing weights for each dictionary atom. The vector $\mathbf{x}$ is estimated by minimizing the distance $dist(\mathbf{y}, \mathbf{Dx})$, where $dist()$ is a distance metric between $\mathbf{y}$ and $\mathbf{Dx}$ such as L_2 norm or Kullback-Leibler (KL)-divergence [33]. In case the dictionary $\mathbf{D}$ is overcomplete, the weight vector $\mathbf{x}$ tends to be sparse. This method of estimating weights

is termed as sparse coding or source recovery. Matching pursuit [34], orthogonal matching pursuit (OMP) [35], basis pursuit [36], focal underdetermined system solver (FOCUSS) [37] and active-set Newton algorithm (ASNA) [33] are some of the source recovery algorithms.

Several methods have been proposed for dictionary learning (DL): random selection of observations [33], K-means clustering [38] , vector quantization [39], dictionary update [40], K-SVD [41], simultaneous codeword optimisation (SimCO) [42] and fast dictionary learning [43]. DL and source recovery methods have been used for classification of objects in images by learning class-specific dictionaries [44]. Shafiee et al. [45] have used three different DL methods to classify faces and digits in images.

The proposed source classification method has been evaluated using ten different noise sources taken from Noisex database [46] and two other instrument sources, one recorded by us and the other, downloaded from an open source portal [47]. The training phase for the audio classification problem is DL from various noise/ instrument sources. We have adopted the recently reported ASNA [33] for source recovery in the testing phase. The advantage of this approach is that the audio sources need not be stationary, since different dictionary atoms capture the variation in the spectral characteristics.

C. Contributions of this work

The main contributions and the novelty of the paper are:

- Dictionary learning by using thresholds on the cosine similarity to ensure distinction amongst the atoms of the same as well as different source dictionaries.
- Proposing two new objective measures, namely, the number of non-zero weights and the sum of weights recovered from ASNA [33] using a concatenation of dictionaries [48], for selecting the most likely audio source from a given set.

II. Proposed method

A. Problem formulation

Given a test audio signal $s[n]$, we need to identify the signal as belonging to one of the audio sources. We train M dictionaries for the M different sources and the test audio signal is classified as that source which gives the highest value for an appropriately defined objective measure.

B. Cosine similarity based dictionary learning

Similar to most of the audio source separation approaches [24], [25], [26], the magnitude of the short-time Fourier transform (mag. STFT) has been used as the feature vector. Feature vectors are L_2 normalized for dictionary learning. Any test feature vector can be represented as an additive, non-negative, linear combination of the dictionary atoms.

After an initialization, each dictionary atom is selected to be as uncorrelated as possible from the rest of the atoms belonging to the same as well as other sources. The correlation between a pair of atoms $\mathbf{d}_n, \mathbf{d}_j$ is measured using the cosine similarity as:

$$cs(\mathbf{d}_n, \mathbf{d}_j) = \mathbf{d}_n^T \mathbf{d}_j / (||\mathbf{d}_n|| ||\mathbf{d}_j||) \quad (1)$$

Two types of cosine similarity measures are used: (a) intra-class cosine similarity (intra-CS) is defined as $cs_i(\mathbf{d}_n, \mathbf{d}_j)$, $\mathbf{d}_n, \mathbf{d}_j \in \mathbf{D}^k, n \neq j$ where $\mathbf{D}^k$ is the dictionary for a specific source; and (b) inter-class cosine similarity (inter-CS) defined as $cs_I(\mathbf{d}_n, \mathbf{d}_j),\ \mathbf{d}_n \in \mathbf{D}^k,\ \mathbf{d}_j \in \mathbf{D}^m, k \neq m$.

For each source, the dictionary atoms are learnt such that the cosine similarity between the atoms is below a set threshold, chosen based on the desired performance. A randomly selected feature vector from first source, denoted as $\mathbf{y}_r$ is taken as the first atom for the first source, $\mathbf{d}_1^1$. The rest of the atoms are learnt by random selection of the feature vectors (excluding features already selected as atoms): t^{th} feature, $\mathbf{y}_t$, is selected as the n^{th} atom, $\mathbf{d}_n^1$ of dictionary $\mathbf{D}^1$ if maximum of intra-CS, $max\ cs_i(\mathbf{y}_t, \mathbf{d}_j^1),\ j < n$ (similar to coherence in [49]) is less than a threshold T_i.

The selection of dictionary atoms is stopped once the number of dictionary atoms reaches a pre-decided number N_A. In case N_A atoms are not obtained, additional mag. STFT features, which do not satisfy the intra-class threshold T_i are appended in the order of increasing $max\ cs_i$.

For learning dictionaries for subsequent sources, atoms are learnt using an additional constraint: $\mathbf{y}_t$ from k^{th} source is selected as the n^{th} atom $\mathbf{d}_n^k$ for the k^{th} dictionary $\mathbf{D}^k$, if $max\ cs_I(\mathbf{y}_t, \mathbf{d}_j^h),\ \mathbf{d}_j^h \in \mathbf{D}^h,\ h < k, 1 \leq j \leq N_A$ is less than a threshold T_I.

The threshold T_i ensures that the atoms within the same source dictionary are as uncorrelated as possible, while T_I ensures that atoms from different source dictionaries are maximally uncorrelated. Lower the values of the thresholds T_i and T_I, greater is the uncorrelatedness between the dictionary atoms.

The total number of atoms in $\mathbf{D}$ from the 12 sources is 1200 using $T_i = T_I = 0.95$ and $N_A = 100$. The proposed DL is summarized in Algorithm 1. For the sake of simplicity, the algorithm does not show the appending of additional dictionary atoms when N_A atoms could not be obtained.

Algorithm 1: Dictionary learning

Initialize: Dictionary index $k = 1$; $\mathbf{D}^1 = \mathbf{d}_1^1 = \mathbf{y}_r$; Atom index $n = 2$; set T_i and T_I.
repeat
  Extract N number of mag.STFT features denoted as $\mathbf{y}_t, 1 \leq t \leq N$ from the k^{th} audio source.
  repeat
    If $n > 1$, find the maximum of intra-CS, m_i as: $\max(cs_i(\mathbf{y}_t, \mathbf{d}_j^k)\ \forall\ j = 1...n-1)$
    If $k > 1$, find the maximum of inter-CS, m_I as: $\max(cs_I(\mathbf{y}_t, \mathbf{d}_j^h)\ \forall\ j = 1..N_A, h < k)$
    if $m_i \leq T_i$ and $m_I \leq T_I$ (for $k > 1$) **then**
      Assign randomly selected $\mathbf{y}_t$ as the n^{th} atom: $\mathbf{d}_n^k = \mathbf{y}_t$ and append to the dictionary: $\mathbf{D}^k = [\mathbf{D}^k\ \mathbf{d}_n^k]$
      $n = n + 1$

10: **end if**
11: **until** $n > N_A$
12: $k = k + 1$; $n = 1$
13: **until** All source dictionaries are learnt
end

C. Metrics for source classification

The learnt dictionaries are used to extract measures for identifying a source. Given an unknown audio signal, the mag. STFT features are extracted, which are used to solve a minimization using ASNA [33]:

$$\underset{\mathbf{x}}{\text{minimize}}\ KL(\mathbf{y}||\hat{\mathbf{y}}),\ \hat{\mathbf{y}} = \mathbf{D}\mathbf{x}\ s.t.\ \mathbf{x} \geq 0 \tag{2}$$

where $KL()$ is the KL divergence between two vectors, $\mathbf{y}$ is the extracted feature, $\hat{\mathbf{y}}$ is the approximation of $\mathbf{y}$, $\mathbf{D}$ is the dictionary using which $\mathbf{y}$ is approximated and $\mathbf{x}$ is the weight vector estimated using ASNA.

Since we know the dictionaries for all the sources, we estimate three measures for classification:

1) *Signal to distortion ratio* (SDR) [50] between $\mathbf{y}$ and $\hat{\mathbf{y}}^i = \mathbf{D}^i\mathbf{x}^i,\ 1 \leq i \leq M$ for the M dictionaries. The SDR with respect to each dictionary $\mathbf{D}^i$ is defined as :

$$SDR^i = 20 \times log_{10}(||\mathbf{y}||_2/||\mathbf{y} - \hat{\mathbf{y}}^i||_2) \tag{3}$$

A feature $\mathbf{y}$ belonging to the k^{th} source can be approximated to a good accuracy by atoms belonging to $\mathbf{D}^k$, since $\mathbf{D}^k$ has been learnt by threshold based selection of atoms from the same source. So, $||\mathbf{y} - \hat{\mathbf{y}}^i||_2$ is expected to be minimum for the k^{th} source, since $\mathbf{y}$ may not be approximated well by atoms from the dictionaries of other sources. Thus, the SDR^i is expected to be maximum for the k^{th} dictionary. The estimated source index $\hat{k}$ for the feature vector of each frame of the test signal is given as $\hat{k} = \arg\max\ SDR^i$.

2) *Number of non-zero weights* (NNZ): We propose this new feature for each source in the weight vector $\mathbf{x}$ recovered using a dictionary $\mathbf{D}$, obtained by concatenating the dictionaries of all the M individual sources: $\mathbf{D} = [\mathbf{D}^1\ \ \mathbf{D}^2 ... \mathbf{D}^M]$. The vector $\mathbf{x} = [\mathbf{x}^{1T}\ \mathbf{x}^{2T} ... \mathbf{x}^{MT}]^T$ obtained by using ASNA on (2) is a concatenation of individual weight vectors $\mathbf{x}^i$ of M sources, and is expected to be sparse.
A test feature vector $\mathbf{y}$ belonging to the k^{th} source can be represented better by atoms from the k^{th} dictionary than by atoms from other dictionaries. Since $\mathbf{D}$ contains atoms from all the sources, the number of non-zero weights, NNZ^k corresponding to the correct dictionary $\mathbf{D}^k$, which is now a sub-matrix of $\mathbf{D}$, may be expected to be higher than $NNZ^i, i \neq k$. The estimated source index $\hat{k}$ for the test vector $\mathbf{y}$ is given by $\hat{k} = \arg\max\ NNZ^i,\ 1 \leq i \leq M$.
The weight vector $\mathbf{x}$ is sparse for the dictionary $\mathbf{D}$, as shown in Fig. 1(a). The number of non-zero weights for each source dictionary is illustrated in Fig. 1(b). For a test frame of babble noise, the highest NNZ is 17 corresponding to babble noise dictionary (atom indices 700 to 800 in $\mathbf{D}$), while 9 is the next highest for the veena dictionary. Thus, a margin of 8 or a factor of 2, is obtained for correct classification.

3) *Sum of weights* (SW) is another scalar measure proposed, defined as the sum of the elements of the vector $\mathbf{x}^i$, which is recovered using the same concatenated dictionary, $\mathbf{D}$. In case the weights are non-sparse, it is observed that SW^i is more reliable than NNZ^i. Figure 1(b) also illustrates the distribution of SW for each of the dictionaries. $\hat{k} = \arg\max\ SW^i$ gives the estimated source index for a test vector $\mathbf{y}$. The sum of weights is the highest (24.47) for babble noise dictionary, while that of veena is 2.33, a factor of about 10.5 for correct classification. It is to be noted that the dictionary used for both NNZ and SW is a concatenated dictionary $\mathbf{D}$, while the measure SDR is derived using separate dictionaries $\mathbf{D}^i$.

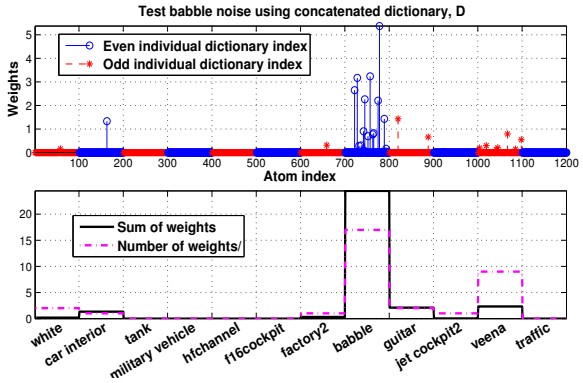

Fig. 1. (a) Weights for a single frame of babble noise estimated by ASNA using concatenated dictionary, $\mathbf{D}$. (b) Number and sum of non-zero weights in (a) as a function of dictionary type for $T_i = T_I = 0.95$.

III. Results and Discussion

Magnitude STFT features are extracted using a frame size of 60 ms and a frame shift of 15 ms from each audio source of duration from 3 to 4 minutes. We experimented with different choices of frame size and arrived at these values as the optimum. Since the number of atoms in each dictionary is constrained to be 100, only 6 seconds from the training set of each audio type form the dictionaries. For evaluating the method, a test signal of duration 5 seconds, equivalent to 330 frames, is taken from the database, and the rest of the audio signal is used in the training stage for learning the dictionaries.

Figure 2 shows the plot of percentage of frames of each test signal correctly classified using SDR as the classification measure for various combinations of T_i and T_I. Table I summarizes the overall audio classification accuracy for different choices of T_i and T_I, where the highest frame level accuracy is obtained for $T_I = T_i = 0.95$ using any of the measures SDR, NNZ and SW. Random selection of mag. STFT features along with the constraint on the cosine similarity has ensured distinct dictionaries and adequate capture of the variations in the audio

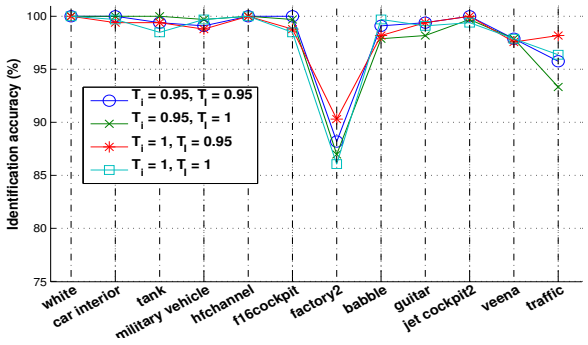

Fig. 2. Percentage of 60 ms frames correctly detected as the original audio source using SDR as the measure, for different choices of T_i and T_I.

TABLE I
OVERALL AUDIO CLASSIFICATION ACCURACY (%) FOR DIFFERENT CHOICES OF T_i AND T_I USING SDR, NNZ AND SW AS MEASURES. MALEH ET. AL. [10]: 89%; GIANNOULIS ET. AL.: 78%

T_i	T_I	SDR	NNZ	SW	MASDR	
					p=6	p=14
0.95	0.95	98.23	87.78	88.51	99.85	100
0.95	1.00	98.01	87.13	88.01	99.64	100
1.00	0.95	98.11	87.05	88.21	99.82	100
1.00	1.00	98.06	87.03	88.42	99.74	99.97

characteristics by the atoms. We have used $T_i = T_I = 0.95$ as the thresholds. Figure 3 shows the percentage of frames correctly classified from each of the 12 audio sources for each of the three measures.

Even though SDR outperforms the other two measures, NNZ and SW are promising since they are computationally simple and give a different insight into the distribution of weights. In case the number of audio sources M is large, using directly SDR as the classification measure is computationally complex, since ASNA is run M number of times. In that case, the measures NNZ or SW can act as the front end for classification (since ASNA is run only once). These measures can pick up the top ranking source dictionaries and then, SDR can be used to find the best fit among them.

A one-vs-one multiclass support vector machine (SVM) classifier is learnt from the training set for the 12 audio classes to compare the performance. It is observed that SVM gives an overall accuracy of 98.1% as compared to 98.23%

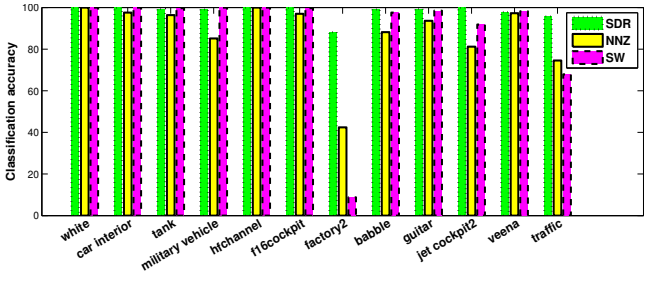

Fig. 3. Individual classification accuracies for all the sources using the three measures independently.

using SDR. Since SDR gives a low accuracy of 88.18% on factory2 noise due to 8.18% of the test frames misclassified as traffic noise, a two class SVM between traffic and factory2 is learnt as secondary classifier. When a frame is classified as traffic, the secondary SVM classifier is used to disambiguate the corresponding frame which increases the classification accuracy of factory2 to 95%. Table II shows the final confusion matrix using SDR measure with $T_i = T_I = 0.95$ and a secondary SVM classifer to resolve the confusion between factory2 and traffic noise. The final classification accuracy over all the audio sources is 98.86%.

In the above discussion we have given frame-wise accuracy. Accuracy can also be computed at the level of a cluster of contiguous frames. Two higher level measures are defined for the i^{th} dictionary, namely, accumulated SDR (ASDR) and moving ASDR (MASDR) as:

$$ASDR_i(q) = \sum_{j=1}^{q} SDR_i(j) \tag{4}$$

$$MASDR_i(q) = \sum_{j=q-p+1}^{q} SDR_i(j) \tag{5}$$

where q is the index of the present frame and p is the number of frames accumulated.

Figure 4 shows the frame-wise SDR and the corresponding ASDR for five test frames of factory noise (worst performing audio source in Fig.2). Only two other audio sources having highest SDR's are shown, for clarity. It is seen in Fig.4 that even though frame-wise SDR for the fourth frame is lower for factory noise, the corresponding ASDR is higher and gives correct classification. In our experiment, we find that 100% classification accuracy can be obtained using MASDR with $p = 6$ for ten of the sources implying that any set of six consecutive frames (135 ms) of the test noise are sufficient for correct classification. Test factory noise requires $p = 10$, and veena, $p = 14$ for 100% classification.

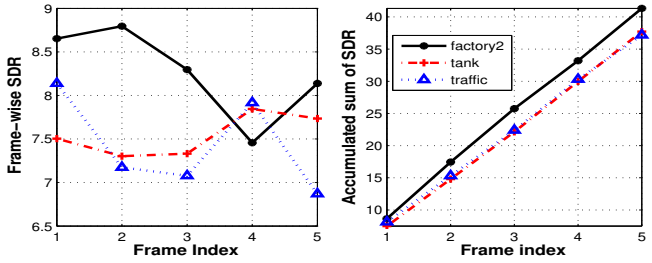

Fig. 4. Advantage of accumulated SDR over frame-wise SDR for $T_i = T_I = 0.95$ for test frames of factory noise.

In addition to showing classification on already known audio source classes, we have shown classification of new noise samples recorded by us in Table III. Similar results on classification of new noises have been shown in Maleh et al. [10]. We have recorded new noise samples in different background environments like bus, mess, railway station,

TABLE II
FINAL CONFUSION MATRIX USING SDR MEASURE WITH $T_i = T_I = 0.95$ AND RESOLVING CONFUSION OF FACTORY2 VS TRAFFIC NOISE USING A SECONDARY SVM CLASSIFIER.

Original/ Estimated	white	car interior	tank	military vehicle	hfchannel	f16cockpit	factory2	babble	guitar	jet cockpit2	veena	traffic
white	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
car interior	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
tank	0.00	0.00	99.39	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.30
military vehicle	0.00	0.00	0.00	99.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.91
hfchannel	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
f16cockpit	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
factory2	0.00	0.91	2.73	0.00	0.00	0.00	95.76	0.00	0.00	0.00	0.00	0.61
babble	0.00	0.00	0.00	0.00	0.00	0.00	0.30	99.09	0.61	0.00	0.00	0.00
guitar	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.39	0.00	0.61	0.00
jet cockpit2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
veena	2.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	97.88	0.00
traffic	0.00	0.00	0.91	0.00	0.00	0.91	0.91	1.21	0.30	0.00	0.00	95.76

market and metro. Given a new noise sample, we mapped each of the frames to one of the already learnt four noise sources using SDR measure. For example, recorded metro noise is classified as hfchannel ($\approx 64\%$), f16 cockpit (22%) and babble noise (14%), which is reasonable since people are speaking in a metro intermittently. This is very useful in the cases where we encounter a new background environment and we need to estimate its composition with respect to already learnt known audio classes.

TABLE III
DISTRIBUTION OF FRAMES (IN %) OF NEWLY RECORDED NOISES CLASSIFIED AS FOUR ALREADY LEARNT NOISE SOURCES I.E. HFCHANNEL, F16COCKPIT, BABBLE AND TRAFFIC USING SDR MEASURE.

Recorded Noise	hfchannel	f16cockpit	babble	traffic
bus	0.00	8.18	89.39	2.42
mess	3.94	31.21	64.85	0.00
railway.stn	29.39	7.27	62.73	0.61
market	17.27	27.58	33.03	22.12
metro	63.64	22.42	13.94	0.00
mall	17.58	0.91	81.52	0.00
construction	43.64	11.52	44.85	0.00

In a real life scenario, the accuracy of classification based on accumulated classification measures is more relevant than individual frame level accuracy, since the classification algorithm gets a stream of test audio signal as input. So, even though a few frames may be individually misclassified, the accumulated classification measure correctly classifies the source.

A. Comparison with previous work

Maleh et. al [10] performed frame-wise noise identification (frame size of 20 ms) using line spectral frequencies as features and pattern recognition based classifiers. They trained using 18.75 minutes of audio data each from 5 noise classes (three of them from NOISEX database), and tested on 500 frames of data for each class. Chu et. al [12] obtained an overall accuracy of 83.9% in recognizing 14 environmental sounds. We have used 12 classes, and obtained an overall frame level accuracy of 98.86% using SDR and a secondary SVM classifier, compared to 89% reported in [10]. The highest accuracy given by majority vote classifier in [22] is around 78%. The accuracy is 100% using MASDR.

IV. CONCLUSION AND FUTURE WORK

A new approach to audio source classification has been proposed adopting ASNA as the source recovery algorithm. Experiments have shown a good overall frame level accuracy of 98.86%. We plan to explore and devise other discriminative dictionary learning and source recovery algorithms for faster and more efficient background source classification. Also, we are working on the classification of the type of background noise from noisy speech and the subsequent separation of speech.

REFERENCES

[1] Yi Hu, and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588-601, Aug. 2007.

[2] D Baby, T Virtanen, and T Barker, " Coupled dictionary training for exemplar-based speech enhancement," *IEEE Int. Conf. Acoust. Speech. Signal Proc. (ICASSP)*, May 2014, pp. 2883 - 2887.

[3] R Turner, [Online] *http://www.wired.co.uk/news/archive/2013-10/02/machine-hearing-cambridge-university*

[4] S. Ikram, "Digital audio forensics using background noise," *Multimedia and Expo (ICME)*, July 2010, pp. 106-110.

[5] R. H. Lyon, "Machinery Noise and Diagnostics," *Butterworth-Heinemann*, 1987.

[6] S. Chu, S. Narayanan, C. C. Jay Kuo, and M. J. Matari, "Where am I? Scene recognition for mobile robots using audio features," *In IEEE International Conference on Multimedia and Expo*, pp. 885-888, 2006.

[7] A. Shirkhodaie, and A. Alkilani, "A survey on acoustic signature recognition and classification techniques for persistent surveillance systems," *Proc. Signal Processing, Sensor Fusion, and Target Recognition*, May 2012.

[8] R. G. Malkin, "Machine listening for context-aware computing," *Doctoral Dissertation, Carnegie Mellon University*, 2006

[9] J. M. Kates, "Classification of background noises for hearing aid applications," *J. Acoust. Soc. Am.*, vol. 91, no.1, pp. 461-470, Jan. 1995.

[10] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame-level noise classification in mobile environments," *Proc. IEEE Conf. Acoustics, Speech, Signal Proc*, March 1999, pp. 237-240.

[11] M. Casey, "Reduced-rank spectra and minimum-entropy priors as consistent and reliable cues for generalized sound recognition," *Proc. Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis, Eurospeech*, Aalborg, Denmark, 2001.

[12] S. Chu, S. Narayanan and C. C. J. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol.17, no.6, 2009.

[13] Z. Liu, J. Huang and Y. Wang, "Classification of TV programs based on audio information using hidden Markov model, *IEEE Signal Processing Society Workshop on Multimedia Signal Processing*, pp. 27-32, Redondo Beach, CA, Dec 1998.

[14] T. Zhang and C. C. Jay-Kuo, "Audio content analysis for online audio visual data segmentation and classification, *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 441-457, May 2001.

[15] L. Lu, H.-J. Zhang and H. Jiang, "Content analysis for audio classification and segmentation, *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 7, pp. 504-516, Oct 2002.

[16] L. Ma, D. J. Smith and B. P. Milner, "Context awareness using environmental noise classification, *Proc. Eurospeech 03*, pp. 2237-2240, Geneva, Switzerland, 2003.

[17] L. Ma, B. Milner and D. Smith, "Acoustic environment classification, *ACM Transactions on Speech and Language Processing*, vol. 3, no. 2, pp. 1-22, July 2006.

[18] C. Couvreura, V. Fontaineb, P. Gaunarda, C. G. Mubikangiey, "Automatic classification of environmental noise events by hidden Markov models, *Proc. ICASSP 98*, 1998.

[19] S. Cherla and V. Ramasubramanian, "Audio analytics by template modeling and 1-pass DP based decoding, *In Proc. INTERSPEECH 10*, pp. 2230-2233, Chiba, Japan, Sep 2010.

[20] V. Ramasubramanian, R. Karthik, S. Thiyagarajan and S. Cherla, "Continuous audio analytics by HMM and Viterbi decoding, *Proc. ICASSP 11*, pp. 2396-2399, Prague, Czech Republic, May 2011.

[21] V. Ramasubramanian, S. Thiyagarajan, G. Pradnya, H. Claussen, J. Rosca, "Two-class verifier framework for audio indexing, *Proc. ICASSP 13*, pp. 838 - 842, Vancouver, Canada, 2013.

[22] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange and M. D. Plumbley, "Detection and classification of acoustic scenes and events: an IEEE AASP challenge," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2013.

[23] B. Cauchi, "Non-negative matrix factorization applied to auditory scene classification," *Masters thesis, ATIAM (UPMC / IRCAM / TELECOM ParisTech)*, 2011.

[24] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol.15, no.3, 2007.

[25] A. Ozerov, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 3, 2010.

[26] G. J. Mysore, P. Smaragdis, and B. Raj, " Non-negative Hidden Markov Modeling of Audio with Application to Source Separation," *Lecture Notes in Computer Science, Latent Variable Analysis and Signal Separation*, vol. 7572, pp. 186-199, 2012.

[27] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 7, 2011.

[28] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," *Interspeech 2010*, Tokyo, Japan, 2010.

[29] Y. C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classication," *Pattern Recognition Letters*, vol. 26, no. 9, 2005.

[30] S. Zubair, F. Yan, W. Wang "Dictionary learning based sparse coefficients for audio classification with max and average pooling," *Elsevier Digital Signal Processing* , vol. 23, issue. 3, 2013.

[31] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: from coding to source separation," *Proceedings of the IEEE,* vol. 98, no. 6, pp. 9951005, 2009.

[32] J. Nikunen and T. Virtanen, "Object-based audio coding using non-negative matrix factorization for the spectrogram representation," *Proceedings of the 128th Audio Engineering Society Convention*, London, UK, 2010.

[33] T. Virtanen, J. F. Gemmeke, B. Raj, "Active-set Newton algorithm for overcomplete non-negative representations of audio", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, pp. 2277 - 2289, 2013.

[34] S. G. Mallat, and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Sig. Process.*, vol. 41, pp. 3397-3415, 1993.

[35] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," *Proceedings of Asilomar Conference on Signals, Systems and Computers*, 1993.

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

[37] I. F. Gorodnitsky, and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted norm minimization algorithm," *IEEE Trans. Sig. Process.*, vol. 45, pp. 600-616, 1997.

[38] A. Coates, and Andrew Y. Ng, "Learning Feature Representations with K-Means," *Lecture Notes in Computer Science, Neural Networks: Tricks of the Trade*, vol. 7700, pp. 561-580, 2012.

[39] K. Kreutz-Delgado, J. Murray, D. Rao, K. Engan, T. Lee, and T. Sejnowski, "Dictionary learning algorithms for sparse representations," *Neural Computation*, vol. 15, pp. 349-396, 2003.

[40] K. Engan, S. O. Aase, and J. H. Husoy, "Multi-frame compression: Theory and design," *EURASIP Signal Process.*, vol. 80, no. 10, pp. 2121-2140, 2000.

[41] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representations," *IEEE Trans. Sig. Process.*, vol. 54, pp. 4311-4322, 2006.

[42] W. Dai, T. Xu, and W. Wang, "Simultaneous Codeword Optimization (SimCO) for dictionary update and learning," *IEEE Trans. Signal Processing*, vol. 60, no. 12, pp. 6340-6353, 2012.

[43] M. G. Jafari, M. D. Plumbley, "Fast dictionary learning for sparse representations of speech signals," *IEEE Journal. Selected Topics Sig. Process.*, vol. 5, pp.1025-1031, 2011.

[44] S. Kong, and D. Wang, " A dictionary learning approach for classification: separating the particularity and the commonality," *Lecture Notes in Computer Science, Computer Vision*, vol. 7572, pp. 186-199, 2012.

[45] S. Shafiee, F. Kamangar, V. Athitsos, and J. Huang, "The role of dictionary learning on sparse representation-based classification," *Proc. Int. Conf. Pervasive Technologies Related to Assistive Environments*, no. 47, 2013.

[46] Noisex-92. [Online], Available: *http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html*

[47] Veena music. [Online], Source: *http://www.youtube.com*

[48] C. Tzagkarakis and A. Mouchtaris, "Sparsity based robust speaker identification using a discriminative dictionary learning approach," *Signal Processing Conference (EUSIPCO)*, 2013, pp. 1–5.

[49] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Info. Theory*, vol. 50, pp. 2231-2242, 2004.

[50] E. Vincent, R. Gribonval, and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, pp. 1462 - 1469, 2006.