2011 IEEE International Workshop on Machine Learning for Signal Processing
September 18-21, 2011, Beijing, China

METHODS FOR LEARNING ADAPTIVE DICTIONARY IN UNDERDETERMINED SPEECH SEPARATION

Tao Xu and Wenwu Wang

Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, United Kingdom
Emails:[t.xu; w.wang]@surrey.ac.uk

ABSTRACT

Underdetermined speech separation is a challenging problem that has been studied extensively in recent years. A promising method to this problem is based on the so-called sparse signal representation. Using this technique, we have recently developed a multi-stage algorithm, where the source signals are recovered using a pre-defined dictionary obtained by e.g. the discrete cosine transform (DCT). In this paper, instead of using the pre-defined dictionary, we present three methods for learning adaptive dictionaries for the reconstruction of source signals, and compare their performance with several state-of-the-art speech separation methods.

Index Terms— Underdetermined blind speech separation, sparse representation, adaptive dictionary learning

1. INTRODUCTION

Over the past two decades, blind source separation (BSS) has attracted a lot of attentions in the signal processing community, owing to its wide range of potential applications, such as in telecommunications, biomedical engineering, and speech enhancement. BSS aims to estimate the unknown sources from their observations without or with little prior knowledge about the channels through which the sources propagate to the sensors. The noise-free instantaneous model of BSS can be described as:

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

where $\mathbf{A} \in R^{M\times N}$ is the unknown mixing matrix assumed to be of full rank, $\mathbf{X} \in R^{M\times T}$ is the observed data matrix whose row vector $\mathbf{x}_i$ is the ith sensor signal having T samples at discrete time instants $t = 1, ..., T$, $\mathbf{S} \in R^{N\times T}$ is the unknown source matrix containing N source vectors.

Many BSS algorithms have been successfully developed for speech separation (or the so-called cocktail party problem), especially for the exactly or over determined cases where the number of microphone mixtures is no smaller than that of the sources. However, for the underdetermined case where $M < N$, it remains an open problem despite many efforts being made in the literature [1].

Underdetermined blind speech separation is an ill-posed inverse problem, due to the lack of sufficient observations. Several approaches have been developed to address this problem. The majority of the approaches [2][3] is based on sparse signal representation. The key idea of sparse signal representation is to assume that the sources are sparse or can be decomposed into the combination of sparse components. Using such a representation, we have recently proposed a novel algorithm [4] based on compressed sensing (CS) in which the BSS model is reformulated to a sparse signal recovery model and extended this approach to a multi-stage model [5] for enhancing the separation performance and improving the computational efficiency. In this paper, we will extend this approach by incorporating an adaptive dictionary learning algorithm for the signal recovery. We will also evaluate the performance of different methods for learning the adaptive dictionaries for the reconstruction of source signals and compare them with the state-of-the-art. The remainder of the paper is organized as follows. The dictionary based underdetermined speech separation approach is presented in Section 2. The three different methods for training the adaptive dictionary are discussed in Section 3. Experimental results are given in Section 4. Finally, conclusions and future work are summarized in Section 5.

2. DICTIONARY BASED UNDERDETERMINED SPEECH SEPARATION

We consider the case of $M = 2$ and $N = 4$, and (1) can be expanded as:

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{pmatrix} \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \mathbf{s}_3 \\ \mathbf{s}_4 \end{pmatrix} \tag{2}$$

where $\mathbf{x}_i (i = 1, 2)$ are the mixtures, $\mathbf{s}_j (j = 1, ..., 4)$ are the sources, and a_{ij} is the ij-th element of the mixing matrix $\mathbf{A}$.

The authors appreciate the funding support from the Engineering and Physical Sciences Research Council (EPSRC) of the UK and China Scholarships Council (CSC).

We can further write the above equation as follows,

$$\underbrace{\begin{pmatrix} x_1(1) \\ \vdots \\ x_1(T) \\ x_2(1) \\ \vdots \\ x_2(T) \end{pmatrix}}_{\mathbf{b}} = \underbrace{\begin{pmatrix} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} & \Lambda_{14} \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} & \Lambda_{24} \end{pmatrix}}_{\mathbf{M}} \underbrace{\begin{pmatrix} s_1(1) \\ \vdots \\ s_1(T) \\ s_2(1) \\ \vdots \\ s_2(T) \\ s_3(1) \\ \vdots \\ s_3(T) \\ s_4(1) \\ \vdots \\ s_4(T) \end{pmatrix}}_{\mathbf{f}} \quad (3)$$

where T is the length of the signal, $\Lambda_{ij} \in R^{T \times T}$ is a diagonal matrix whose diagonal elements are all equal to a_{ij}. Let $\mathbf{b} = vec(\mathbf{X}^T)$, $\mathbf{f} = vec(\mathbf{S}^T)$, where vec is an operator stacking the column vectors of a matrix into a single vector. Equation (3) can be written in a compact form as:

$$\mathbf{b} = \mathbf{M}\mathbf{f} \quad (4)$$

The above equation can be interpreted as a signal recovery problem in a CS model, in which $\mathbf{M}$ is the measurement matrix, which can be estimated by a clustering algorithm, e.g. [5] and $\mathbf{b}$ is the compressed vector of samples in $\mathbf{f}$. Moreover, a sparse representation in the transform domain can be employed for $\mathbf{f}$:

$$\mathbf{f} = \mathbf{\Phi}\mathbf{y} \quad (5)$$

where $\mathbf{\Phi}$ is a transform dictionary and $\mathbf{y}$ contains the weighting coefficients in the $\mathbf{\Phi}$ domain. Combining (4) and (5), we have

$$\mathbf{b} = \mathbf{M}\mathbf{\Phi}\mathbf{y} \quad (6)$$

According to compressed sensing theories, if both $\mathbf{M}$ and $\mathbf{\Phi}$ satisfy certain conditions [6] and also $\mathbf{y}$ is sparse, the signal $\mathbf{f}$ can be recovered by the measurement $\mathbf{b}$ using an optimization process.

Therefore, designing the dictionary $\mathbf{\Phi}$ is an important issue for the signal recovery algorithms. According to recent research, two main approaches are usually used: the analytical approach and the learning-based approach. In the first approach, a mathematical model of the data is given in advance so that the dictionary can be generated by fast Fourier transform (FFT), discrete cosine transform (DCT), wavelet transform, etc. The second approach applies machine learning techniques to train the dictionary from a set of data so that the dictionary atoms obtained represent the feature of the signal. Dictionary learning methods are often established on an optimization algorithm, in which an initial dictionary is given and a signal is decomposed as a linear combination of the atoms from the initial dictionary and the weighted values are a few non-zero coefficients. Then the atoms of the dictionary are trained when the weighting coefficients are fixed. After that, the trained dictionary is used to compute the new weighting coefficients. The process is iterated until the most suitable dictionary is learned eventually [7] [8] [9], based on a pre-defined criterion.

Finally, Figure 1 is the flow chart summarizing the main processes in the dictionary based underdetermined speech separation system described above, where the same method as in [5] has been used in the clustering and separating stages. The part of dashed box represents dictionary designing step to be discussed in the following section.

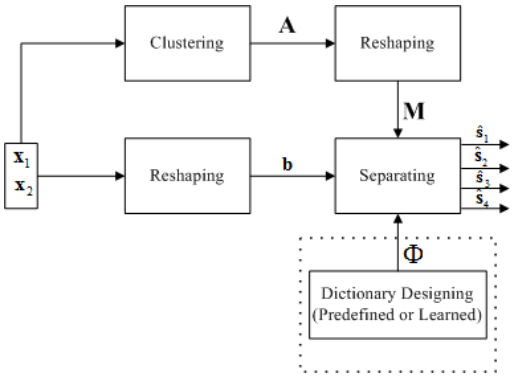
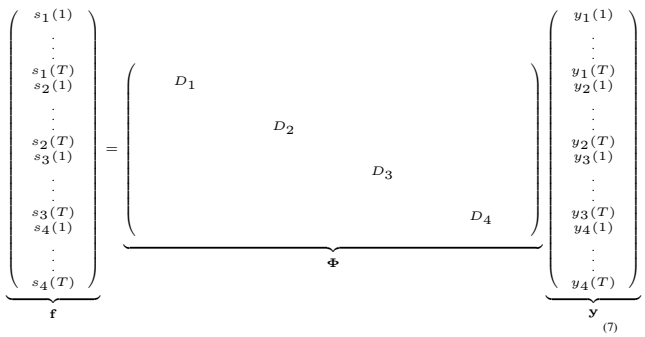


Fig. 1. The flow chart of the proposed system for separating four speech sources from two mixtures.

3. THE STRATEGIES FOR TRAINING THE ADAPTIVE DICTIONARY

In our previous work [5], the pre-defined DCT dictionary was used in the dashed box of Figure 1 to decompose the signal based on the first approach described in the above section. Experimental results reveal that it is potentially beneficial to design a more suitable dictionary using the second approach, i.e. learning an adaptive dictionary to capture the feature of the signal. In this work, the K-SVD algorithm [7] is used to obtain the dictionary atoms in our proposed system. The K-SVD algorithm aims to iteratively find the best dictionary to represent the signal samples. It consists of a sparse-coding step and a dictionary updating step. The first step is to compute the sparse coefficient vectors from the sample signals using any sparse-approximation approach such as a pursuit method based on the given dictionary. The second step is to update the atoms which are the columns in the dictionary matrix to better fit the signal using the sparse representations obtained in the first step. The dictionary update is carried out for one atom each time, while keeping the other atoms fixed. These two steps are iteratively repeated until the convergence of the algorithm.

However, how to train the dictionary is an important practical issue. Here we suggest three different training strategies. To this end, we first expand (5) as follows.



(7)

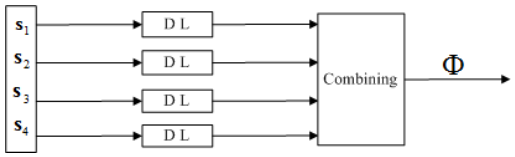


Fig. 2. The flow chart of the STD strategy.

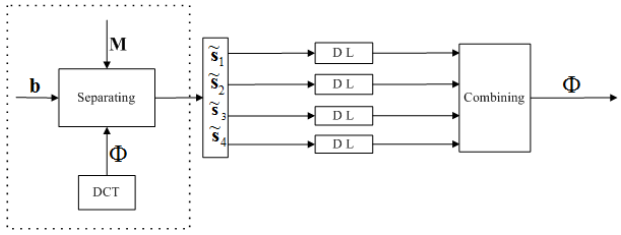


Fig. 3. The flow chart of the ESTD strategy.

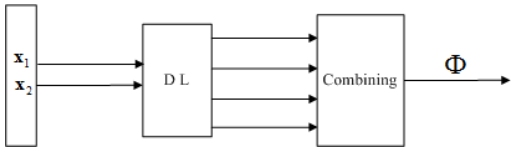


Fig. 4. The flow chart of the MTD strategy.

The first strategy called the source-trained dictionary (STD) is depicted in Figure 2 where DL represents dictionary learning. In this method, for each source, we train a dictionary. Therefore four different dictionaries D_1, D_2, D_3, D_4 are trained from the four original sources respectively. They are then combined to form a single dictionary matrix $\mathbf{\Phi}$ for separating the source in the following stages. For example, the D_1 in Equation (7) is trained from the source $\mathbf{s}_1$. Firstly, the speech source vector is reshaped to a speech sample matrix which contains consecutive speech frames (each frame has L samples) from the source vector with an overlap of P samples (to ensure a sufficient number of signals in the sample matrix). Therefore, the sample matrix has L rows and $\lfloor (T-L)/(L-P) \rfloor + 1$ columns, where $\lfloor . \rfloor$ rounds the argument to its nearest integer. Then the dictionary is computed by the K-SVD algorithm as a $L \times L$ matrix. Finally, the dictionary matrix is arranged in diagonal form with an overlap of $L/2$ samples until the $T \times T$ dictionary D_1 is filled in. By using this block diagonal operation we essentially split the signal in small vectors and there will typically be a block boundary issue, i.e. a jump between the joining area of two adjacent blocks, causing undesired artifacts in the coefficients.

To avoid this we can multiply the vector by a window function (e.g. the Hamming window), thus smoothing the signal at the boundaries. Some information may be lost because of the windowing, and hence overlapping between blocks is used to eliminate this problem. In our experiments, 50% overlap between the vectors is used. The other dictionaries D_2, D_3, D_4 can be generated in the same way. The final single dictionary matrix $\mathbf{\Phi}$ is formed by arranging these four dictionaries along the diagonal of $\mathbf{\Phi}$ without overlaps. Ideally, the order of the dictionaries $D_i (i = 1, ..., 4)$ should be consistent with the order of sources $\mathbf{s}_i$. According to our experiments, when a mismatch of the orders occurs, the separation performance may be degraded. The reason that this happens could be that the feature of a speech source is better captured by its corresponding dictionary rather than the dictionary obtained from another source.

The second strategy called the estimated source-trained dictionary (ESTD) is depicted in Figure 3. Firstly the sources are estimated from the mixtures by the algorithm [5] using e.g. the DCT. Secondly, the dictionaries D_1, D_2, D_3, D_4 are learned from these four coarsely separated sources, whose atoms are then used to reconstruct the sources. These four dictionaries are used to form the dictionary matrix $\mathbf{\Phi}$. Each estimated source $\tilde{\mathbf{s}}_i (i = 1, ..., 4)$ used to train the dictionary is segmented with an overlap of P samples (each frame has L samples) to form the sample matrix. The learned dictionaries $D_i (i = 1, ..., 4)$ are obtained from this sample matrix in the same way as described in STD by using the K-SVD algorithm.

The third strategy, namely the mixture-trained dictionary (MTD), is illustrated in Figure 4. The two mixtures $\mathbf{x}_i (i = 1, 2)$ used to train the dictionary are segmented with an overlap of P samples (each frame has L samples) to form the sample matrix which has L rows and $(\lfloor (T-L)/(L-P) \rfloor + 1) * 2$ columns. Then the dictionary is computed by the K-SVD algorithm as a $L \times L$ matrix. Finally, the dictionary is arranged in diagonal form with an overlap of $L/2$ samples until the $T \times T$ dictionary D_M is filled in. In this method, D_1, D_2, D_3, D_4 in Equation (7) are all identical to D_M which is trained from the mixtures by the same methods as used for the above two strategies.

In comparison, as shown in the next section, the STD has the best performance among the three methods. This suggests that the dictionary trained in this way best matches the original speech source. However, this approach requires the sources to be available *a priori* when training the dictionary. Although in BSS, the sources are assumed to be unknown, the STD method shows the potential performance that can be achieved by a dictionary learning approach. The MTD estimates the sources in a blind manner, as it trains the dictionary directly from mixtures. Nevertheless it captures the features less accurately from each source as compared with STD. The ESTD method provides a good trade-off. On the one hand, it estimates the sources automatically without the availability of the sources for dictionary learning as required in STD method. On the other hand, it offers separation performance better than the MTD method.

4. EXPERIMENTAL RESULTS

In this section, we evaluate the proposed algorithm by performing the experiments using four speech sources in the TIMIT database, which are English male (EM), English female (EF), Japanese female (JF) and Chinese female (CF) speech respectively. The sources have a duration of 5 seconds, sampled at 10 kHz. For objective quality assessment, we use the two global performance criteria defined in the BSSEVAL toolbox [10] to evaluate the performance from the estimated source signals, which are the signal to distortion ratio (SDR) and the source to interference ratio (SIR), defined respectively as

$$SDR = 10log_{10}\frac{\parallel s_{target} \parallel^2}{\parallel e_{interf} + e_{noise} + e_{artif} \parallel^2} \quad (8)$$

$$SIR = 10log_{10}\frac{\parallel s_{target} \parallel^2}{\parallel e_{interf} \parallel^2} \quad (9)$$

where $s_{target}(t)$ is an allowed deformation of the target source $s_i(t)$, $e_{interf}(t)$ is an allowed deformation of the sources which accounts for the interferences of the unwanted sources, $e_{noise}(t)$ is an allowed deformation of the perturbation noise (but not the sources), and $e_{artif}(t)$ is an artifact term that may correspond to artifacts of the separation algorithm such as musical noise. Therefore, the estimated source $\hat{s}(t)$ can be decomposed as follows:

$$\hat{s}(t) = s_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t) \quad (10)$$

The parameters L and P are set to 512 and 450 samples respectively. The parameter controlling the sparsity of the coefficient vector in K-SVD was set to 5. The mixtures are generated by the mixing matrix used in [5]. From the mixtures, we can recover the four speech sources using the DCT dictionary and the adaptive dictionaries based on the STD, ESTD and MTD methods. The results are presented in Table 1 and 2.

	DCT	STD	ESTD	MTD
EM speech	7.59	9.89	6.93	-1.41
EF speech	9.53	11.44	9.26	3.54
JF speech	2.73	7.38	2.13	-4.22
CF speech	14.59	15.05	14.13	8.91

Table 1. SDR (in dB) measured for each estimated speech source.

	DCT	STD	ESTD	MTD
EM speech	14.23	19.47	14.49	2.81
EF speech	11.35	30.49	11.35	5.25
JF speech	6.07	22.21	5.97	-2.12
CF speech	18.12	25.89	18.34	12.50

Table 2. SIR (in dB) measured for each estimated speech source.

From these tables, we can observe that the separation performance using STD trained dictionary is considerably better than using the DCT dictioary. Using the ESTD trained dictionary, the results are close to the DCT dictionary. However, it is difficult to obtain good results by using the dictionary learned from the mixtures, i.e. the MTD method. These results suggest that the properly learned dictionaries outperform the pre-defined dictionary in underdetermined speech separation.

We also compared the proposed algorithm with other methods i.e. algorithm 2 [11] in the campaign SiSEC2010 [12] and algorithms 4 [13], 7 [14] and 11 [15] in the campaign SiSEC2008 [16]. For the comparison of these algorithms, the same speech sources, mixtures and the same evaluation method and dataset are used in this experiment. The dataset contains four female speech sources having a duration of 10 seconds, sampled at 8kHz. That is, each signal has 80000 samples.

In comparison with other algorithms involved in the campaign (see also the results reported in [16]), the proposed algorithm provides competitive separation performance. To show this, we plot the results measured by the SDR in Figure 5 and the SIR in Figure 6 for each source tested in the experiments. For SDR, the proposed algorithm with STD obtains better performance than the algorithm 11 for three speech sources. For SIR, the proposed algorithm with STD training scheme obtains better results than all of other algorithms for speech source 2 and 3. The average performance obtained by the STD method over the four sources is considerably better than all the compared methods, while the results obtained by ESTD and MTD are comparable to these methods. In our informal subjective listening tests, the proposed algorithm was also observed to offer good separation quality.

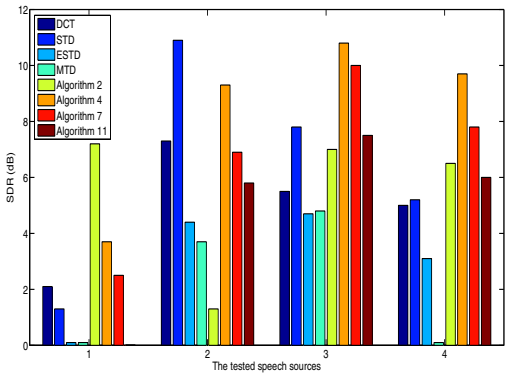


Fig. 5. SDR comparison between the proposed algorithm and four recent algorithms (i.e. algorithm 2, 4, 7 and 11 in the campaigns), with sources 1-4 corresponding to the four female speech signals 1-4 in the campaign.

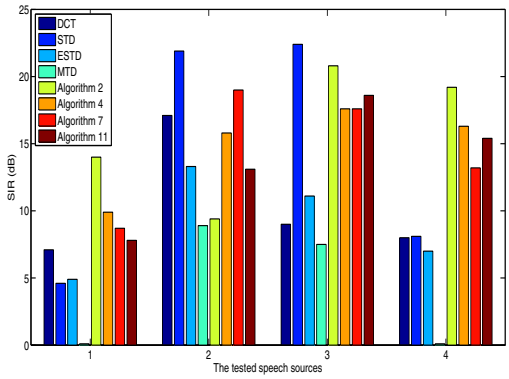


Fig. 6. SIR comparison between the proposed algorithm and four recent algorithms (i.e. algorithm 2, 4, 7 and 11 in the campaigns), with sources 1-4 corresponding to the four female speech signals 1-4 in the campaign.

5. CONCLUSIONS AND FUTURE WORK

We have presented a promising method to underdetermined speech separation based on sparse representation with adaptive dictionary learning. In particular, we have suggested three different training methods for obtaining the dictionary matrix. Numerical experiments have shown the competitive separation performance by the proposed method, as compared with several underdetermined BSS approaches reported in the recent source separation evaluation campaigns. This study has also shown that it is advantageous to use an adaptive dictionary as compared to a pre-defined dictionary. In the future, we will improve the dictionary learning algorithm to achieve better separation performance.

6. REFERENCES

[1] S. Makino, T. W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, 2007.

[2] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, Nov. 2001.

[3] R. Gribonval and S. Lesage, "A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges," in *European Symposium on Artificial Neural Networks*, 2006, pp. 323–330.

[4] T. Xu and W. Wang, "A compressed sensing approach for underdetermined blind audio source separation with sparse representation," in *Proc. IEEE Int. Workshop on Statistical Signal Processing*, 2009, pp. 493 – 496.

[5] T. Xu and W. Wang, "A block-based compressed sensing method for underdetermined blind speech separation incorporating binary mask," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 2022 – 2025.

[6] E. Candes and T. Tao, "The dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, pp. 2313–2351, 2007.

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

[8] R. Rubinstein, S. Member, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," in *IEEE Trans. Signal Processing*, 2010, vol. 58, pp. 1553 – 1564.

[9] M. G. Jafari and M. D. Plumbley, "Dictionary learning for speech based on a doubly sparse greedy adaptive dictionary algorithm," preprint, 2010.

[10] E. Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[11] S. Rickard, *The DUET Blind Source Separation Algorithm*, Springer Netherlands, 2007.

[12] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. J. Theis, G. Nolte, D. Lutter, and N. Q. K. Duong, "The 2010 signal separation evaluation campaign (sisec2010): audio source separation," in *Proc. LVA/ICA*, 2010, pp. 114–122.

[13] E. Vincent, S. Arberet, and Rémi Gribonval, "Underdetermined instantaneous audio source separation via local gaussian modeling," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation*, 2009, pp. 775–782.

[14] Y. Deville, M. Puigt, and B. Albouy, "Time-frequency blind signal separation: extended methods, performance evaluation for speech sources," in *Proc. IEEE International Conference on Neural Networks*, 2004, pp. 255–260.

[15] B. V. Gowreesunker and A. H. Tewfik, "Blind source separation using monochannel overcomplete dictionaries," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 33–36.

[16] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. ICA*, 2009, pp. 734–741.