MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Deep clustering: Discriminative embeddings for segmentation and separation

Hershey, J.R.; Chen, Z.; Le Roux, J.; Watanabe, S.

TR2016-003 March 2016

Abstract

We address the problem of "cocktail-party" source separation in a deep learning framework called deep clustering. Previous deep network approaches to separation have shown promising performance in scenarios with a fixed number of sources, each belonging to a distinct signal class, such as speech and noise. However, for arbitrary source classes and number, "class-based" methods are not suitable. Instead, we train a deep network to assign contrastive embedding vectors to each time-frequency region of the spectrogram in order to implicitly predict the segmentation labels of the target spectrogram from the input mixtures. This yields a deep network-based analogue to spectral clustering, in that the embeddings form a low-rank pairwise affinity matrix that approximates the ideal affinity matrix, while enabling much faster performance. At test time, the clustering step "decodes" the segmentation implicit in the embeddings by optimizing K-means with respect to the unknown assignments. Preliminary experiments on single channel mixtures from multiple speakers show that a speaker-independent model trained on two-speaker mixtures can improve signal quality for mixtures of held-out speakers by an average of 6dB. More dramatically, the same model does surprisingly well with three-speaker mixtures.

2016 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2016
201 Broadway, Cambridge, Massachusetts 02139

DEEP CLUSTERING: DISCRIMINATIVE EMBEDDINGS FOR SEGMENTATION AND SEPARATION

John R. Hershey[1]*, Zhuo Chen*[2]*, Jonathan Le Roux*[1]*, Shinji Watanabe*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA
[2]Columbia University, New York, NY, USA

ABSTRACT

We address the problem of "cocktail-party" source separation in a deep learning framework called *deep clustering*. Previous deep network approaches to separation have shown promising performance in scenarios with a fixed number of sources, each belonging to a distinct signal class, such as speech and noise. However, for arbitrary source classes and number, "class-based" methods are not suitable. Instead, we train a deep network to assign contrastive embedding vectors to each time-frequency region of the spectrogram in order to implicitly predict the segmentation labels of the target spectrogram from the input mixtures. This yields a deep network-based analogue to spectral clustering, in that the embeddings form a low-rank pairwise affinity matrix that approximates the ideal affinity matrix, while enabling much faster performance. At test time, the clustering step "decodes" the segmentation implicit in the embeddings by optimizing K-means with respect to the unknown assignments. Preliminary experiments on single-channel mixtures from multiple speakers show that a speaker-independent model trained on two-speaker mixtures can improve signal quality for mixtures of held-out speakers by an average of 6dB. More dramatically, the same model does surprisingly well with three-speaker mixtures.

***Index Terms*—** speech separation, embedding, deep learning, clustering

1. INTRODUCTION

In real world perception, we often must selectively attend to objects whose features are intermingled in the incoming sensory signal. Nowhere is this more apparent than in hearing, where signals are densely mixed and can be challenging to separate. Nevertheless human listeners easily perceive separate sources in an acoustic mixture, and this ability has inspired a variety of computational approaches to the so-called *auditory scene analysis* or *cocktail party* problem [1]. We address the problem of "cocktail-party" speech separation in a deep learning framework we call *deep clustering*.

Single-channel speech separation is the task of estimating the individual speech signals that are mixed together and overlapping in a monaural signal. It is a challenging problem and many further assumptions have been used to make headway. Previous attempts have generally assumed that the number of sources is fixed. Some "speech separation" methods are about separating speech from challenging background noise [2–4] instead of separating multiple speakers. Many previous approaches have relied on speaker-dependent models [5–7], although some also addressed the case of same-speaker mixtures [5, 7], or more than two speakers [8]. Furthermore, many of these addressed only tasks with limited vocabulary and grammar, as in [9]. Some were able to achieve impressive performance in these limited domains.

In this work, we consider a more open and difficult task of speaker-independent separation of two or more speakers, with no special constraint on vocabulary and grammar. Speaker-independent separation was addressed in [10] by building speaker adaptation upon the model-based approach of [5]. In another direction, [8] extended [5] to handle more than two speakers. While both extensions are interesting, in general speed and learning are problematic.

Meanwhile, the state-of the art in enhancement and separation is currently done using deep networks [11–13]. These *class-based* methods train on parallel sets of mixtures and their constituent target sources, so that the network predicts the source belonging to the target class, or classifies the type of source that dominates each time-frequency bin.

Although *class-based* methods can succeed in the speaker-dependent case, where each target is a speaker known at training time, they fail to learn in the speaker-independent case, as shown in our experiments. Neural networks have an output dimension for each target source class, and when targets are multiple sources of the same type, they encounter a *permutation problem*: the system needs to make an arbitrary decision about which output dimension to use for each source. For neural networks, which deterministically map a given input to a given source estimate in each dimension, it is difficult to learn if the the permutation of training targets into slots is arbitrary and indeterminate.

An important family of methods based on clustering may be more flexible in this regard. These include *computational auditory scene analysis* (CASA) approaches that use perceptual grouping cues [14, 15], and spectral clustering approaches [16] that use affinity kernels. CASA approaches seek to explain perceptual grouping of regions in terms of their similarity [17]. Such methods are heuristic, and although carefully tuned systems perform surprisingly well on speech [18], they still fall behind the class-based deep learning methods, as we show below. With no training, over-fitting is not a problem, but it is difficult to imagine accommodating different types of sources.

In the area of spectral clustering, however, which is based on eigen-decomposition of the normalized affinity matrix [19], significant progress has been made in learning the relative weights of different affinity features [16]. Unfortunately, the spectral clustering paradigm suffers from high computational cost, and shallow learning. These factors appear to be co-dependent: simple kernels tend to produce sparse affinity matrices, which require costly spectral methods to reduce to clusters. Conversely this very complexity makes optimization of the front end processing a formidable challenge [16].

In the speech separation problem, powerful front-end processing is indeed required, because of a pesky chicken and egg problem. To infer the segmentation requires features of neighboring regions of

This work was done while Z. Chen was an intern at MERL.

the same source, but the context regions for one source also contain intermingled parts of other sources. To extract uncorrupted features, then, would seem to require knowing the segmentation in advance.

Nevertheless, we know from from prior work that deep neural networks can learn their way out of this quandary, when the targets are distinct classes. So we propose to use more powerful front end processing to produce a lower-rank affinity matrix, which then may be amenable to clustering by simpler methods such as K-means. The simpler clustering methods in turn should provide for easier training, allowing a more complex front-end to be learned.

Learned feature transformations known as *embeddings* have recently been gaining significant interest in many fields. Unsupervised embeddings obtained by auto-associative deep networks, used with relatively simple clustering algorithms, have recently been shown to outperform spectral clustering methods [20, 21] in some cases.

In our framework a deep network assigns embedding vectors to each time-frequency region of the spectrogram, according to an objective function that minimizes the distances between embeddings of time-frequency bins dominated by the same source, while maximizing the distances between embeddings for those dominated by different sources. Thus the clusters in the embedding can represent the inferred spectral masking patterns of the sources, in a permutation-free way. Moreover, despite the fixed dimensionality of the network output, the embeddings can implicitly represent different numbers of sources.

This objective relates to spectral clustering in that the embeddings can be used to approximate an ideal affinity matrix given by the known segmentation. It is also closely related to the K-means objective function so that at test time we can infer the assignments given the embeddings using K-means algorithm.

The experiments show that the proposed method can separate speech using a speaker-independent model on an open set of speakers. We derive partition labels by mixing signals together and observing their spectral dominance patterns. After training on a database of mixtures of speakers trained in this way, we show that the model can generalize to three-speaker mixtures despite training only on two-speaker mixtures. Although results are preliminary, this suggests that we may hope to achieve class-independent segmentation of arbitrary sounds, with additional application to image segmentation and other domains.

2. LEARNING DEEP EMBEDDINGS FOR CLUSTERING

We define as x a raw input signal and as $X_i = g_i(x), i \in \{1, \ldots, N\}$, a feature vector indexed by an element i. In the case of audio signals, i is typically a time-frequency index (t, f), where t indexes frame of the signal, f indexes frequency, and $X_i = X_{t,f}$ the value of the complex spectrogram at the corresponding time-frequency bin. We assume that there exists a reasonable partition of the elements i into regions, which we would like to find, for example to further process the features X_i separately for each region. In the case of audio source separation, these regions can be defined as the sets of time-frequency bins in which each source dominates, and estimating such a partition would enable us to build time-frequency masks to be applied to X_i, leading to time-frequency representations that can be inverted to obtain isolated sources.

To estimate the partition, we seek a D-dimensional embedding $V = f_\theta(x) \in \mathbb{R}^{N \times D}$, parameterized by θ, such that performing some simple clustering in the embedding space will likely lead to a partition of $\{1, \ldots, N\}$ that is close to the target. In this work, $V = f_\theta(X)$ is based on a deep neural network that is a global function of the entire input signal X. Thus our transformation can take into account global properties of the input, and the embedding can be considered a permutation- and cardinality-independent encoding of the network's estimate of the signal partition. Here we consider a unit-norm embedding, so that $|v_i|^2 = 1$ where $v_i = \{v_{i,d}\}$ and $v_{i,d}$ is the value of the d-th dimension of the embedding for element i. We consider the embeddings V to implicitly represent an $N \times N$ estimated affinity matrix VV^T.

The target partition is represented by the indicator $Y = \{y_{i,c}\}$, mapping each element i to each of C clusters, so that $y_{i,c} = 1$ if element i is in cluster c. In this case YY^T, is considered as a binary affinity matrix that represents the cluster assignments in a permutation-independent way: $(YY^T)_{i,j} = 1$ if elements i and j belong to the same cluster, and $(YY^T)_{i,j} = 0$ otherwise, and $(YP)(YP)^T = YY^T$ for any permutation matrix P.

We can learn affinity matrix VV^T, as a function of the inputs, X to match the affinities, YY^T, by minimizing, with respect to $V = f_\theta(X)$, the training cost function,

$$\mathcal{C}_Y(V) = \|VV^T - YY^T\|_{\mathrm{F}}^2 = \sum_{i,j} \left(\langle v_i, v_j\rangle - \langle y_i, y_j\rangle\right)^2 \quad (1)$$

$$= \sum_{i,j:y_i=y_j} \left(|v_i - v_j|^2 - 1\right) + \sum_{i,j} \langle v_i, v_j\rangle^2, \quad (2)$$

summed over training examples, where $\|A\|_{\mathrm{F}}^2$ is the squared Frobenius norm. For the true cluster labels $\mathring{Y}$, $\mathcal{C}_{\mathring{Y}}(V)$ minimizes the distance between the estimated affinity matrix VV^T and the ideal affinity matrix $\mathring{Y}\mathring{Y}^T$. The form (2) pulls the embeddings v_i and v_j closer together for elements within the same partition, whereas the second term pushes all elements apart, preventing collapse to a trivial solution.

Note that although this function ostensibly sums over all pairs of data points i, j, the low-rank nature of the objective leads to an efficient implementation:

$$\mathcal{C}_Y(V) = \|V^TV\|_{\mathrm{F}}^2 - 2\|V^TY\|_{\mathrm{F}}^2 + \|Y^TY\|_{\mathrm{F}}^2, \quad (3)$$

which avoids explicitly constructing the $N \times N$ affinity matrix. In practice, N is orders of magnitude greater than D, leading to a significant speedup. Derivatives with respect to V are also efficiently obtained due to the low-rank structure:

$$\frac{\partial \mathcal{C}_Y(V)}{\partial V^T} = 4V(V^TV) - 4Y(Y^TV) \quad (4)$$

This low-rank formulation also relates to spectral clustering in that the latter typically requires the Nyström low-rank approximation to the affinity matrix [22] for efficiency. So, rather than making a low-rank approximation to a complicated full-rank model, deep clustering directly optimizes a low-rank model so that simple clustering can be used.

For inference, we compute the embeddings $V = f_\theta(X)$ on the test signal X, and cluster the rows $v_i \in \mathbb{R}^D$, by minimizing the K-means inference cost: $\bar{Y} = \arg\min_Y \mathcal{K}_V(Y) = \|V - YM\|_{\mathrm{F}}^2$, where $M = (Y^TY)^{-1}Y^TV$ are the $C \times D$ means of each cluster. The resulting cluster assignments $\bar{Y}$ are used as binary masks to separate the sources. The ideal mask used as our cluster reference $\mathring{Y}$, yields the optimal signal to noise ratio (SNR) among all binary masks. Although continuous masks can yield further improvement, here we first focus on solving the permutation problem, leaving refinement for future work.

The clustering error between the estimates $\bar{Y}$, and the labels $\mathring{Y}$, can be quantified as in [23] using

$$d(\bar{Y}, \mathring{Y}) = \|\bar{Y}(\bar{Y}^T\bar{Y})^{-1}\bar{Y}^T - \mathring{Y}(\mathring{Y}^T\mathring{Y})^{-1}\mathring{Y}^T\|_{\mathrm{F}}^2. \quad (5)$$

The function $\mathcal{C}_{\mathring{Y}}(V)$ can be shown to bound the clustering error on the same data:

$$d(\bar{Y}, \mathring{Y}) \leq \eta \sqrt{\mathcal{C}_{\mathring{Y}}(V)}, \tag{6}$$

where $\eta = 4\sqrt{C(D+C)}/\min_c(N_c)$, with N_c the size of cluster c. Thus reducing the training objective on some data brings the optimal K-means clustering of that data closer to the reference clustering. There are a variety of alternative training and inference objective functions, along with their associated error bounds, and we leave their exploration for future work.

3. SPEECH SEPARATION EXPERIMENTS

3.1. Experimental setup

We evaluate deep clustering (DC) on a speaker-independent speech separation task. Mixtures involving speech from same gender speakers can be extremely challenging since the pitch and vocal tract of the voices are in the same range. We here consider mixtures of two and three speakers, which include the same gender condition. Three types of experiments were performed, separating two unknown speakers, three unknown speakers, or three known speakers. In the latter case, the systems are trained on mixtures of the three known speakers at training time, whereas in the other cases training speakers and test speakers are different.

We created a new corpus of speech mixtures using utterances from the Wall Street Journal (WSJ0) corpus because existing speech separation challenge datasets are too limited for the evaluation of our model. For example, the speech separation challenge [9] only contains two-speaker mixtures, with a limited vocabulary and insufficient training data.

A 30 h training set and a 10 h validation set consisting of two-speaker mixtures were generated by randomly selecting utterances by different speakers from the WSJ0 training set `si_tr_s`, and mixing them at various signal-to-noise ratios (SNR) between 0 dB and 10 dB. The validation set was used to optimize some tuning parameters and to evaluate the source separation performance in closed conditions (**CC**). Five hours of evaluation data were generated similarly using utterances from 16 speakers from the WSJ0 development set `si_dt_05` and evaluation set `si_et_05`. The speakers are different from those in our training and validation sets, and we thus use this set for open condition (**OC**) evaluation. Note that previous speech separation methods (e.g., [24, 25]) cannot handle the open speaker problem, and require knowledge of the speakers in the evaluation.

We also created three sets of three-speaker mixtures. The first two sets are similar respectively to the two-speaker validation and evaluation sets, with 100 three-speaker mixtures obtained from a pool of many speakers in closed condition (**MS-CC**) and open condition (**MS-OC**). The third one consists in 5000 mixtures for training, 500 mixtures for validation, and 500 mixtures for test, using speech from a closed set of three known speakers in `si_et_05` (**3S-CC**) .

All data were downsampled to 8 kHz before processing to reduce computational and memory costs. The input features X were the log spectral magnitudes of the speech mixture, computed using a short-time Fourier transform (STFT) with 32 ms window length, 8 ms window shift, and the square root of the hann window. To ensure local coherency, a mixture is separately processed in half-overlapping segments of 100 frames, roughly the length of one word in speech, to output embeddings V based on the proposed model.

3.2. Training procedure

The binary masks were used to build the target Y to train our network. In each time-frequency bin, the mask values are set to 1 for the source with the maximum magnitude and 0 for the others. For the two-source case, this corresponds to the ideal binary mask (IBM) [26]. To avoid training the network to assign embeddings to silence regions, a binary weight for each time-frequency bin was used during the training process, only retaining those bins such that magnitude of the mixture at that bin is greater than some ratio (arbitrarily set to -40 dB) of the maximum magnitude. The network structure used in our experiments has two bi-directional long short-term memory (BLSTM) layers, followed by one feedforward layer. Each BLSTM layer has 600 hidden cells and the feedforward layer corresponds with the embedding dimension D. Stochastic gradient descent with momentum 0.9 and fixed learning rate 10^{-5} was used for training. In each updating step, to avoid local optima, Gaussian noise with zero mean and 0.6 variance was added to the weight. We prepared several networks using different embedding dimensions from 5 to 60. In addition, two different activation functions (logistic and tanh) were explored to form the embedding V with different ranges for $v_{n,d}$. For each embedding dimension, the weights for the corresponding network were initialized randomly according to a normal distribution with zero mean and 0.1 variance with the tanh activation. In the experiments with the logistic activation, the network was initialized with the tanh network.

A state of the art class-based BLSTM speech enhancement network [11] was included as baseline for both two-speaker and three-speaker experiments. Because of the inherent ambiguity in speaker-independent separation tasks, as to which output should be used for each speaker, we proposed two training schemes to help with learning using the class-based LSTM. In one case we used the stronger source as the training target for each 100 frame segment (**BLSTM stronger**). We also propose a permutation-free scheme (**BLSTM permute**), where we find the closest clean source to each output of the network, and use that source to measure the training error and compute the gradients.

To facilitate comparison, both deep clustering and the classifier system used the same architectures, except for the final output layers and objective function. Since deep clustering has a large embedding layer, we also formulated a class-based BLSTM with the same number of parameters by using an additional feedforward layer of the same size as the embedding layer used in deep clustering (**BLSTM permute***). In the three-known-speakers experiment, the speaker identities are known, so we used the stacked ideal soft mask for each speaker as target (**BLSTM stack**). For both experiments, squared Euclidean distance was used as error measurement for class-based network. All the BLSTM layers in the class-based model were initialized with the parameters of the trained deep clustering network (i.e. $D = 40$ tanh).

3.3. Speech separation procedure

At test time, speech separation was performed by re-filtering time-domain signals based on time-frequency masks for each speaker. The masks were obtained by clustering the row vectors of embedding V, where V was output from the proposed model for each segment (100 frames), similarly to the training stage. The number of clusters is set to the number of speakers in the mixture. We evaluated two types of clustering methods: global K-means on the embeddings of the whole utterance and local K-means, where clustering is done separately on each 100-frame segment. In both cases, we choose the

Table 1: SDR improvements (dB) for different separation methods

method	CC	OC
oracle NMF	5.1	-
CASA	2.9	3.1
DC local K-means	6.5	6.5
DC global K-means	5.9	5.8
BLSTM stronger	1.3	1.2
BLSTM permute	1.3	1.3
BLSTM permute*	1.4	1.2

Table 2: SDR improvements (dB) for different embedding dimensions D and activation functions

	CC		OC	
model	DC local	DC global	DC local	DC global
$D = 5$	-0.8	-1.0	-0.7	-1.1
$D = 10$	5.2	4.5	5.3	4.6
$D = 20$	6.3	5.6	6.4	5.7
$D = 40$	6.5	5.9	6.5	5.8
$D = 60$	6.0	5.2	6.1	5.3
$D = 40$ logistic	6.6	5.9	6.6	6.0

best correspondence in the least-squares sense between the recovered sources and target signals.

Given that DC can represent an arbitrary number of clusters, an interesting question is whether it can generalize to the case of three-speaker mixtures without changing the model parameters. Speech separation experiments on three-speaker mixtures were thus conducted using the network trained with two-speaker mixtures, by simply changing the number of clusters from 2 to 3 in the clustering step.

Besides the class-based BLSTM, we used supervised sparse non-negative matrix factorization (SNMF) as another baseline [24, 25]. While SNMF is amenable to separating male-female mixtures when using a concatenation of bases trained on speakers of different genders, in preliminary experiments it failed for same-gender mixtures. We thus give SNMF an unfair advantage by using speaker dependent models with oracle information about the speakers present at test time. Wiener-filter like masks are built using the estimated models and applied to the mixture, and the separated signals are obtained by inverse STFT. We used 256 bases per speaker, and magnitude spectra with 8 consecutive frames of left context as input features. We also included an unsupervised CASA-based system [18] as another baseline for the two-speaker experiment.

For all experiments, performance was evaluated in terms of averaged signal-to-distortion ratio (SDR) using the `bss_eval` toolbox [27]. The initial SDR averaged over the mixtures was 0.2 dB for two-speaker mixtures and -3.0 dB for three-speaker mixtures.

4. RESULTS AND DISCUSSION

As shown in Table 1, both local and global clustering methods significantly outperform all baselines. Note that due to stability issues with the CASA code provided by authors of [18], evaluation could only be run on a subset of about 40 % of the data, but there was no significant difference for this subset in starting SNR or in the improvements of other algorithms. The global K-means clustering of the whole utterance performs only slightly worse than local clustering. As the system was only trained with individual segments, this suggests that the network learns globally important features. The performance of DC is similar in open and closed conditions, indicating that it can generalize well to unknown speakers.

In Table 2, the $D = 5$ system completely fails, either because optimization of the current network architecture fails, or the embedding fundamentally requires more dimensions. The performance of $D = 20$, $D = 40$, $D = 60$ is similar, showing that the system can operate in a wide range of parameter values. We arbitrarily used tanh networks in most of the experiments because of their larger embedding space than logistic networks. However, in Table 2, we verify that the logistic network performs about the same.

Table 3: SDR improvement (dB) for mixtures of three speakers. Left: three-speaker separation using DC network trained on two-speaker mixtures. Right: separation of three known speakers.

method	MS-CC	MS-OC
oracle NMF	4.4	-
DC local	3.5	2.8
DC global	2.7	2.2

method	3S-CC
oracle NMF	4.5
DC local	7.0
DC global	6.9
BLSTM stack	6.8

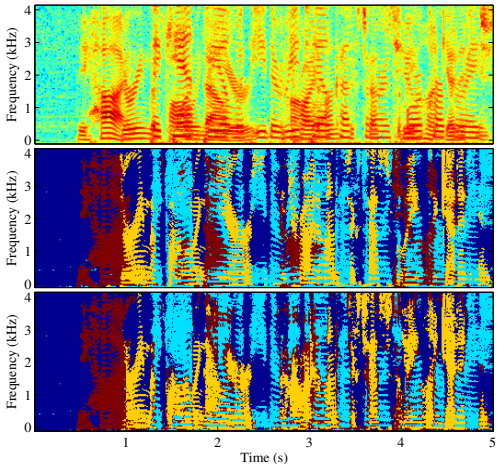

Fig. 1: Example of three-speaker separation. Top: mixture log spectrogram. Middle: IBM. Dark blue shows silence. Bottom: output mask from proposed system trained on two-speaker mixtures.

All class-based BLSTMs performed poorly in non-speaker-dependent settings, even when carefully trained (Table 1, right). Only for the speaker-dependent 3S-CC set, the class-based model performed similarly to DC (Table 3). We can expect other speaker-dependent methods [6, 28] to follow the same trend. This confirms that class-based networks lack the ability to resolve the permutation problem introduced by same-class mixtures. In contrast, in DC the permutation is solved by the clustering step, which allows modeling power to focus on the distinction between sources.

We see in Table 3 (left) that DC remarkably can also separate three-speaker mixtures, even when only trained on two-speaker mixtures. Figure 1 shows an example of separation for three-speaker mixture in the open validation set. Of course, including mixtures involving more than two speakers at training time should improve performance further, but the method does surprisingly well even without retraining. Performance is now worse than oracle NMF, but is again much better once we allow DC to focus on a limited set of speakers, as shown in Table 3 (right): there, DC is trained on mixtures of the same three speakers used for test.

We evaluated deep clustering in a variety of conditions and parameter regimes, on a challenging speech separation problem. Since these are preliminary results, we expect further refinement of the model will lead to significant improvements. Alternative network architectures with different time and frequency dependencies, such as deep convolutional neural networks [29] or hierarchical recursive embedding networks [30], could be helpful in terms of learning and regularization. Finally, scaling up training on databases of more disparate audio types, as well as applications to other domains such as image segmentation, are prime candidates for future work.

5. REFERENCES

[1] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1990.

[2] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014.

[3] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 7, 2013.

[4] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 2, 2013.

[5] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Comput. Speech Lang.*, vol. 24, no. 1, 2010.

[6] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *arXiv preprint arXiv:1502.04149*, 2015.

[7] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space," in *Proc. Interspeech 2006*, Pittsburgh, 2006.

[8] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition," *IEEE Signal Process. Mag.*, vol. 27, no. 6, 2010.

[9] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, 2010.

[10] R. J. Weiss, "Underdetermined source separation using speaker subspace models," Ph.D. dissertation, Columbia University, 2009.

[11] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Latent Variable Analysis and Signal Separation*. Springer, 2015.

[12] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 12, 2014.

[13] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *Signal Processing Letters, IEEE*, vol. 21, no. 1, 2014.

[14] M. P. Cooke, "Modelling auditory processing and organisation," Ph.D. dissertation, Univ. of Sheffield, 1991.

[15] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, MIT, 1996.

[16] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *JMLR*, vol. 7, 2006.

[17] M. Wertheimer, "Laws of organization in perceptual forms," in *A Source book of Gestalt psychology*, W. A. Ellis, Ed. Routledge and Kegan Paul, 1938.

[18] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, 2013.

[19] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. PAMI*, vol. 22, no. 8, 2000.

[20] F. Tian, B. Gao, Q. Cui, E. Chen, and T.-Y. Liu, "Learning deep representations for graph clustering," in *Proc. AAAI*, 2014.

[21] P. Huang, Y. Huang, W. Wang, and L. Wang, "Deep embedding network for clustering," in *Proc. ICPR*, 2014.

[22] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nyström method," *IEEE Trans. PAMI*, vol. 26, no. 2, 2004.

[23] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, 1985.

[24] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, 2007.

[25] J. Le Roux, F. J. Weninger, and J. R. Hershey, "Sparse NMF – half-baked or well done?" MERL, Cambridge, MA, USA, Tech. Rep. TR2015-023, Mar. 2015.

[26] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005.

[27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, 2006.

[28] K. Hu and D. Wang, "An iterative model-based approach to cochannel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, 2013.

[29] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. PAMI*, vol. 35, no. 8, 2013.

[30] A. Sharma, O. Tuzel, and M.-Y. Liu, "Recursive context propagation network for semantic scene labeling," in *Proc. NIPS*, 2014.