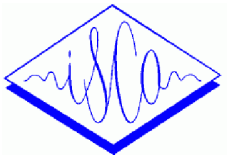

Third Workshop on Spoken Language
Technologies for Under-resourced Languages
Cape Town, South Africa
May 7-9, 2012

SUBSPACE-GMM ACOUSTIC MODELS FOR UNDER-RESOURCED LANGUAGES: FEASIBILITY STUDY

Xueru Zhang, Kris Demuynck, Dirk Van Compernolle, Hugo Van hamme*

KU Leuven, Department of Electrical Engineering - ESAT
Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven, Belgium
{Xueru.Zhang, Kris.Demuynck, Dirk.VanCompernolle, Hugo.Vanhamme}@esat.kuleuven.be

ABSTRACT

Acoustic model parameter estimation is hampered by a lack of data. To reduce the number of parameters to be estimated, we propose *sub-GMM* modelling, which constrains the acoustic models to a low-dimensional manifold embedded in the space of Gaussian mixture weights. The manifold model is obtained through non-negative matrix factorization with sparsity constraints. Our preliminary monolingual experiments show that the proposed model is as efficient as clustering the distributions to a smaller set, while it opens perspectives for a new parameter tying technique. In the example, the number of parameters to be estimated per distribution is reduced more than an order of magnitude.

Index Terms— under-resourced languages, manifold, sparsity, non-negative matrix factorization, substructure

1. INTRODUCTION

State-of-the art context dependent (CD) hidden Markov model (HMM) based speech recognition systems model each CD phone by a multi-state HMM. Nowadays, acoustic models are trained on hundreds to thousands of hours of speech [1], where we observe an effect of diminishing returns as we add more data. Even for this size of the training data, it is not feasible to reliably estimate the large numbers of parameters in CD models without parameter tying techniques, as will be discussed below. For so-called *under-resourced* languages, the amount of training data that is available is only a few hours to tens of hours, which motivates this effort to revisit parameter tying. The amount of data aligned to each CD phone state is unevenly distributed. For some HMM states, there are sufficient training data to estimate a Gaussian mixture model (GMM) with a sufficient number of components to obtain a good accuracy. For other states there may be very little or even no data to reliably estimate the hundreds or thousands of parameters of a state GMM. To mitigate this scarce data problem, GMM parameters are tied with various methods and at different hierarchical levels. For instance, complete phones may be shared (customary for e.g. context independent filled pause model), state emission density models (GMMs) can be shared in a phonetic decision tree (PDT) [2][3], or Gaussian components can be shared among different GMMs, a method that is also applied in this study with algorithmic and optimisation aspects discussed in [4]. In the PDT, CD states are clustered for similarity of their GMMs and eventually share a common GMM. Because the criteria for traversing the decision tree are phonetic questions about the conditioning phone context, they have the property of generalizing to unseen contexts. While a PDT drastically reduces the number of GMMs to be trained, it needs to be created from language-specific data. A phonetic decision tree derived from another language cannot be cloned to the under-resourced language, given that the tree leaves which are populated in the other language may not be populated for the under-resourced language, while other leaves may be over-represented. Several techniques have been proposed to cope with the data scarcity problem when handling under-resourced languages. Bootstrapping [5], adaptation and cloning models from existing languages [6] have shown to improve the performance of speech recognition systems when only limited amounts of training data are available. [7] explores the combination of universal feature detectors with language specific rule-based phone models. In [8], a rapid language adaptation ASR system for under-resourced languages based on multilingual unsupervised training is described. The cross-language transfer hypotheses are used as transcriptions.

In this paper, we exploit the fact that the distributions of speech features cannot take any form and hence must be constrained to a manifold. We further propose a linear model on the mixture weights in GMMs to describe the speech manifold. Non-negative matrix factorization (NMF) [9] is used to estimate the manifold model from trained GMMs. The fundamental idea of NMF is to approximate a non-negative matrix as a product of two non-negative matrices: a low-rank matrix of *latent vectors* and the latent vector coefficient matrix. The latent vectors can be interpreted as the GMM weights of *sub-GMMs*, i.e. the NMF decomposes every GMM into a weighted addition of these sub-GMMs.

The primary goal of this paper is to evaluate the feasibility of this new parameter tying concept that exploits sparsity properties of non-negative linear combinations. While it is not a finished end-to-end study, this paper describes the idea and tests its fundamental hypothesis that the GMMs can be adequately represented by a restricted set of sub-GMMs that are combined *sparsely*. Also, it reports on the algorithms to construct the sub-GMMs.

This paper is organized as follows. In section 2, the geometric formulation of the low-dimensional manifold is illustrated. In section 3, we explain the manifold model. In section 4, NMF and the resulting modelling with mixtures of sub-GMMs is introduced. We describe the sparsity constraints in detail in section 5. In section 6 we describe our speech recognition system and analyze the recognition results. Conclusions are presented in section 7.

2. GEOMETRIC PERSPECTIVE

Given the few degrees of freedoms of the speech generation apparatus, the human voice cannot produce any arbitrary sound. When

*This work is funded by the Dutch-Flemish IMPact program (ICTRegie-IBBT(Interdisciplinary Institute for Broadband Technology)) and by FWO grant G.A122.10N AMODA.

short speech segments are mapped to an acoustic space, the reachable points lie on a manifold which is embedded in the high dimensional acoustic space [10]. This low-dimensional manifold is presumed to be general and to persist across languages. It is reasonable to assume that this manifold can be described by making acoustic models for all sounds in a set of reference languages. Any sound of a new, possibly under-resourced, language lies on this manifold and should have a coordinate on the manifold.

In this paper, the acoustic space is parameterised by the mixture weights of Gaussians shared across all CD states (the last tying mechanism described in section 1). A point in this acoustic space represents a GMM. In [4], it was shown that, at least for monolingual data, this parametrisation with shared Gaussians can result in accurate acoustic models. In the proposal of this paper, the pool of shared Gaussians is envisaged to be obtained from training HMMs on a set of reference languages, but the actual method for doing so is not the focal point of this paper. The set of reference languages should be rich enough to *span* all sounds of the target language in the sense that will become clear below.

Speech recognition models are described by the GMMs of all its CD states. Hence, a language is described by a collection of GMMs that occupy a set of points in the GMM space. All languages together define a manifold of interest in the GMM space. For languages with abundant data, the GMM parameters can be estimated reliably, hence providing *reference* points on the speech manifold. For an under-resourced language it is not possible to measure points on the manifold. Instead, a manifold model is used to constrain the GMMs so they can be estimated reliably from small amounts of data. In other words, by exploiting the manifold constraint, only the coordinates on this low-dimensional manifold need to be estimated.

3. MANIFOLD MODEL

Let N be the number of Gaussians, shared among all states. The GMM for HMM state s is completely described by the N mixture weights λ_{ns}, which lie on the $(N-1)$-simplex $\sum_{n=1}^{N}\lambda_{ns}=1$ with $\lambda_{ns}\geq 0$, in which the speech manifold is embedded. Different methods can be used to constrain a point (a GMM) to the manifold. In *quantization*, a noisy estimate (due to lack of data) of a GMM is replaced by the closest (e.g. in the sense of Kullback-Leibler divergence) reference point on the manifold. In other words, each GMM of the new language is quantized by its unique closest neighbour in one of the reference languages. Hence, a nearest-neighbour (vector quantization), memory or exemplar representation of the manifold is applied.

In the present paper, we propose a linear model to describe the manifold.

$$\lambda_{\mathbf{s}}=\mathbf{W}\mathbf{h}_s \tag{1}$$

where $\lambda_{\mathbf{s}}=[\lambda_{1s},\cdots,\lambda_{Ns}]^t$ is the set of GMM weights that model CD phone state s. $\mathbf{W}$ is the $N\times L$ matrix containing the non-negative mixture weights of the $L<N$ sub-GMMs and $\mathbf{h}_s$ is the state-specific vector of non-negative weights with which the sub-GMMs are linearly combined. The columns of $\mathbf{W}$ and the vector $\mathbf{h}_s$ are constrained to unity L_1 norm, so eqn. (1) constrains the GMM λ to a $(L-1)$-dimensional subspace. However, with non-negativity and sparsity constraints on $\mathbf{h}_s$, the GMM models are constrained further. The manifold model $\mathbf{W}$ is estimated from a total of S observed reference points:

$$\mathbf{V}\approx\mathbf{W}\mathbf{H} \tag{2}$$

with $\mathbf{V}_{N\times S}$, $\mathbf{W}_{N\times L}$ and $\mathbf{H}_{L\times S}$. where $\mathbf{V}$ is the matrix obtained by stacking all reference points (GMM coefficient vectors) as its columns. Other than in the quantization model described above, the manifold model $\mathbf{W}$ is a regression model through the reference points and is hence more robust to estimation uncertainty in the GMMs of the reference language. Also, eqn. (1) allows sparse interpolation between the sub-GMMs (columns of $\mathbf{W}$).

The model finally contains L sub-GMMs, modelling the speech manifold linearly. This is very different from a traditional *clustering* approach where the S reference GMMs would be assigned to one of L clusters. Each cluster is a GMM *joining* its member GMMs, thus spreading the probability mass over a larger area in feature space. Contrarily, the proposed model seeks to obtain GMMs with a smaller extent, hence the name "sub-GMMs".

4. NON-NEGATIVE MATRIX FACTORIZATION

In this research, we use non-negative matrix factorization (NMF) [11] to estimate the Gaussian mixture weights of the sub-GMMs needed to describe a new language. Non-negative matrix factorization, which performs matrix factorization and dimension reduction, approximates a non-negative matrix $\mathbf{V}$ as a product of a lower rank non-negative latent vector matrix $\mathbf{W}$ and a corresponding non-negative coefficient matrix $\mathbf{H}$ which are compact and rich enough to approximately reconstruct the original matrix. The matrix factors are found by minimizing a distortion measure between $\mathbf{V}$ and $\mathbf{WH}$, such as Kullback- Leibler divergence (KLD) [9], Euclidean distance (EU), Itakura-Saito divergence or the more general formulations with α-divergence, β-divergence, γ-divergence [12]. The choice of cost function depends on the underlying probabilistic formulation. As will be motivated below, a maximum likelihood formulation will lead to the KLD cost function. The KLD between the mixture weights of two GMMs also approximates the KLD between the GMMs provided the overlap between the Gaussians is small, which makes it a natural method to measure how well a GMM is approximated by a point on the manifold.

The matrices $\mathbf{W}$ and $\mathbf{H}$ in eqn. (2) are chosen to maximize the likelihood of the training data from all the reference languages, which is equivalent to maximizing the following auxiliary function:

$$Q(\mathbf{W},\mathbf{H})=\sum_{n,s}\gamma_{ns}\log(\lambda_{ns}) \tag{3}$$

where γ_{ns} is the accumulated posterior probability of Gaussian n at HMM state s. The Gaussian mixture weights λ_{ns} are constrained to a linear combination of L latent vectors $\mathbf{w}_l$ which span the manifold given by eqn. (1). To satisfy the normalization constraint of λ_{ns}, it is required that

$$\left\{\begin{array}{lcl}\sum_n w_{nl} &=& 1,\forall l\\ \sum_l h_{ls} &=& 1,\forall s\end{array}\right. \tag{4}$$

The KLD cost function for NMF is given by

$$D(\mathbf{V}||\mathbf{WH})=\sum_{n,s}\left[v_{ns}\log\frac{v_{ns}}{(\mathbf{WH})_{ns}}-v_{ns}+(\mathbf{WH})_{ns}\right] \tag{5}$$

When choosing $[\mathbf{V}]_{ns}=\gamma_{ns}$, minimizing eqn. (5) over $\mathbf{W}$ and $\mathbf{H}$ under constraints eqn. (4) is equivalent to maximizing eqn. (3) under the same constraints.

The resulting multiplicative update rule for $\mathbf{W}$ and $\mathbf{H}$ are similar to the ones given in [9]:

$$\left\{\begin{array}{lcl}w_{nl} &=& \frac{1}{\eta_l^W}w_{nl}\sum_s\frac{\gamma_{ns}h_{ls}}{(\mathbf{WH})_{ns}}\\ h_{ls} &=& \frac{1}{\eta_s^H}h_{ls}\sum_n\frac{\gamma_{ns}w_{nl}}{(\mathbf{WH})_{ns}}\end{array}\right. \tag{6}$$

with η_l^W and η_s^H the normalization factors, which assure that the constraints from eqn. (4) are satisfied. The fixed point updates of $\mathbf{W}$ and $\mathbf{H}$ are repeated for a fixed number of times (see section 6) which is assumed to be sufficient for convergence.

For a new under-resourced language and given the estimated $\mathbf{W}$, the coordinates on the manifold can be estimated by maximizing the likelihood of the language-specific speech data. Maximizing the likelihood is equivalent to maximizing the auxiliary function (3) as a function of $\hat{\mathbf{h}}$. The multiplicative update formula for $\hat{\mathbf{h}}$ is

$$\hat{h}_l = \hat{h}_l \sum_{n=1}^{N} \frac{\gamma_n w_{nl}}{\sum_{j=1}^{L} w_{nj} \hat{h}_j} \tag{7}$$

where γ_n is the posterior probability of Gaussian n accumulated over the available data.

5. SPARSITY

Both $\mathbf{W}$ and $\mathbf{H}$ are expected to have sparse structures. Each column of $\mathbf{W}$ should form a sub-GMMs: each GMM is a linear combination of these substructures, i.e. the columns $\mathbf{W}$ are the *parts* or *atoms* that the GMM's are composed of. It is to be expected that these parts contain less Gaussians than the original GMMs themselves, i.e. $\mathbf{W}$ should be more sparse than $\mathbf{V}$. If it would not be more sparse, it would rather be a clustering of the GMMs into less accurate, merged units and NMF would not be finding the substructures of the GMMs. On the other hand, we don't want the trivial decomposition that the sub-GMMs are the Gaussians themselves i.e. a column of $\mathbf{W}$ has only one non-zero element.

To generate a sparse $\mathbf{W}$ and meanwhile prevent the uninformative matrix factorization case, extra constraints are applied on both $\mathbf{W}$ and $\mathbf{H}$. By applying these sparsity constraints, the columns of $\mathbf{W}$ will be constrained to localize towards the substructures of the GMMs. Sparsity is often implemented by L_1-regularization [13]. However, while these methods can increase the sparsity, they do not enforce a maximum on the L_0 norm of each column, i.e. the number of non-zero elements (this is also the case for [14]). Hence, we opted for an alternative approach: during the iteration eqn. (6), sparsity in $\mathbf{W}$ and $\mathbf{H}$ are enforced by counting the number of non-zeros elements in each column of $\mathbf{W}$ and $\mathbf{H}$ and if the counts are larger than a target K_W and K_H respectively, the smallest entries are set to zero. Given the zero-locking property of update eqn. (6), elements will remain zero once set to zero. Meanwhile, the columns with less than K_W or K_H non-zero elements should allow for additional non-zeros. In other words, Gaussian components should be allowed to migrate from sub-GMMs with too many components to sub-GMMs with too few components.

The algorithm for non-negative matrix decomposition with maximum column support (the number of non-zero elements) of K_W and K_H of respectively $\mathbf{W}$ and $\mathbf{H}$ is outlined in 1. In other research, we have obtained better NMF results with a simulated annealing technique [15] where in each iteration of eqn. (6), we add noise matrices homomorphous to $\mathbf{W}$ resp. $\mathbf{H}$. The noise matrices are constructed with i.i.d. entries uniformly distributed between 0 and 1, normalized column-wise to unit L_1-norm, scaled with $\mathbf{W}$ resp. $\mathbf{H}$ and with an exponentially decreasing function of the iteration number. With this technique, the cost function values obtained are generally smaller, i.e. it is better at avoiding local extrema. By setting $K_H = L$ (or $K_W = N$) no sparsity is enforced on $\mathbf{H}$ (or $\mathbf{W}$).

Algorithm 1 Non-negative matrix decomposition with maximum column support of K_W resp. K_H.

Step 1 : Set the $N \times L$ mask matrix $\mathbf{M}_W$ equal to all ones. Set the $L \times S$ mask matrix $\mathbf{M}_H$ equal to all ones.
Step 2 : Perform an NMF without sparsity constraints using simulated annealing. In each iteration, multiply $\mathbf{W}$ with $\mathbf{M}_W$ and multiply $\mathbf{H}$ with $\mathbf{M}_H$.
Step 3 : For each column of $\mathbf{W}$, count the number of non-zero elements. If larger than K_W, set the entries in $\mathbf{M}_W$ corresponding to the 10% smallest entries of the column to zero (rounded to the nearest integer and at least one). Likewise, for each column of $\mathbf{H}$, count the number of non-zero elements. If larger than K_H, set the entries in $\mathbf{M}_H$ corresponding to the 10% smallest entries of the column to zero.
Step 4 : If all columns of $\mathbf{W}$ have at most K_W non-zero elements, and all columns of $\mathbf{H}$ have at most K_H non-zero elements, stop; else goto step 2.

6. EXPERIMENTS

6.1. Experimental setup

As mentioned in the introduction, the main goal of this paper is to evaluate the sub-GMM model. A cross-language experiment will require at least an HMM training in a set of reference languages, a method for building a compact Gaussian set (i.e. Gaussians are shared efficiently across languages) and also to develop a method to construct a phonetic tree from scarce data. Note that the maximum likelihood formulation of section 4 offers opportunities for clustering based on the sparse weights h_{ls}. These approaches will require further in-depth studies and are beyond our current focus. In the current experiments, we want to evaluate whether the proposed manifold model is capable of producing HMMs with sufficient accuracy. Rather than clustering GMMs (sparsity in $\mathbf{W}$ smaller than sparsity in $\mathbf{V}$ – see section 2) the NMF model should yield decomposition into sub-HMMs (sparsity in $\mathbf{W}$ greater than sparsity in $\mathbf{V}$).

Training is done on the SI-284 data from the Wall Street Journal (WSJ) comprising 81 hours from 284 speakers. The baseline speech recognizer used in our experiments is a semi-tied Gaussian mixture HMM system. The system uses a shared pool of 32754 Gaussians to model the observations in 5967 cross-word context-dependent tied triphone states (GMMs). All acoustic units –context-dependent variants of one of the 42 phones or silence– have a 3-state left-to-right topology. The silence HMM is excluded from this analysis. The acoustic features consist of 22 MEL spectra with mean normalization and VTLN, augmented with their first and second order time derivatives, which results in 66 dimensional feature vectors. These features are then mapped to a 39 dimensional space by means of a discriminative linear transformation and decorrelation.

For developing and evaluating the system, we combined the WSJ 5k closed and 20k open vocabulary non-verbalized punctuation Nov92 and Nov93 tasks. By combining all evaluation data, we obtained one large evaluation set containing 101 minutes of speech (18298 words). The combination of all corresponding development data was used to tune system parameters such as pruning thresholds and the weight ratio between the language model and the acoustic model.

The proposed speech manifold model is used to decompose and reconstruct the Gaussian mixture weights of the acoustic models. The performance of the proposed NMF method with 1800 iterations for eqn. (6) is evaluated by applying the reconstructed Gaussian mix-

HMM	L	K_W	$\mathbf{V}_{0.99}$	$\mathbf{W}_{0.99}$	$\mathbf{H}_{0.99}$	WER(%)
SI_{cd}	5964	/	84.7	/	/	6.42
	1007		167.6			7.19
sub-GMM	1000	N	84.7	132.9	2.6	7.18
	2000			104.0	2.9	6.74
	2500			100.5	2.8	6.67
	3000			97.1	2.7	6.39
	1000	100	84.7	81.8	5.2	7.10
		80		67.9	6.5	7.28

Table 1. Word error rate. SI_{cd} : CD speaker independent baseline system. The L for SI_{cd} refers to the number of tied states.

ture weights $\hat{\lambda} = \mathbf{WH}$ as the new Gaussian mixture weights of the acoustic models.

6.2. Results

As explained in section 3, it is key that the sub-GMM models use less Gaussians that the original GMMs, i.e. that the columns of $\mathbf{W}$ are more sparse than those of $\mathbf{V}$. The sparsity metric used here, called *99% sparsity*, is the average (over columns) of the cardinality of the minimal set of elements required to cover 99% of the probability mass of a column. This is listed in Table 1 as $\mathbf{W}_{0.99}$ and $\mathbf{H}_{0.99}$. For comparison, the trivial decomposition with $L = 5964$ and $\mathbf{W} = \mathbf{V}$ and $\mathbf{H}$ equal to the identity matrix is shown. For comparison with the case $L = 1000$, a model where the PDT was pruned back to 1007 densities is also built (second row in Table 1. The models without sparsity constraints do not generate sub-GMMs, as their $\mathbf{W}_{0.99}$ are larger than 84.7. Rather than modelling a subspace, these models have a *clustering* effect, which is also seen from their low value of $\mathbf{H}_{0.99}$. The models with sparsity constraint do succeed in reducing $\mathbf{W}_{0.99}$, especially when compared to the model with the PDT with 1007 leaves.

The word error rates (WER) in Table 1 show that the NMF without sparsity constraints seems to be as accurate in clustering as pruning back the PDT. More importantly, with sparsity constraints, the accuracy loss is minor or non-existent, making the sub-GMM model a usable manifold parametrisation.

On average, 6.5 non-zero weights in $\mathbf{H}$ are needed to model the source GMMs. This number of parameters is sufficiently low to be estimated from scarce training data. But the manifold model allows to generate a greater variety of distribution than quantizing to the nearest GMM in $\mathbf{V}$. Since the number of coefficients in $\mathbf{h}$ to be estimated is already small, we did not observe obvious performance differences in our preliminary experiments by also limiting K_H.

7. CONCLUSIONS AND FUTURE RESEARCH

In this paper, a new concept to tie model parameters for under-resourced languages is explained. A low-dimensional manifold is described by a linear model. By applying NMF with sparsity constraints on $\mathbf{W}$ (and $\mathbf{H}$), meaningful sub-GMMs are discovered. The number of parameters to be estimated per GMM is reduced more than an order of magnitude. This method is quite appealing for training a phonetic decision tree for under-resourced languages by only estimating the speech data points on the manifold (i.e. $\mathbf{h}$-coordinates) in each node of the tree, with increasing the likelihood as the splitting criterion.

The proposed method is applied to monolingual data. The obvious next step is to proceed as outlined in the first paragraph of section 6.

8. REFERENCES

[1] R. Prasad, S. Matsoukas, C.-L. Kao, J. Ma, D.-X. Xu, T. Colthurst, O. Kimball, R. Schwartz, J. Gauvain, L. Lamel, H. Schwenk, G. Adda, and F. Lefevre, "The 2004 BBN/LIMSI 20xRT english conversational telephone speech recognition system," September 2005, pp. 1645–1648.

[2] M.-Y. Hwang, X. Huang, and F. Alleva, "Predicting unseen triphones with senones," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 6, pp. 412–419, 1996.

[3] W. Reichl and W. Chou, "Robust decision tree state tying for continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 555–566, 2000.

[4] K. Demuynck, *Extracting, Modelling and Combining Information in Speech Recognition*, Ph.D. thesis, Katholieke Universiteit Leuven, 2001.

[5] V.-B. Le and L. Besacier, "Automatic speech recognition for under-resourced languages: Application to vietnamese language," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 8, pp. 1471–1482, 2009.

[6] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.

[7] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Toward a detector-based universal phone recognizer," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4261–4264.

[8] N. T. Vu, F. Kraus, and T. Schultz, "Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training," August 2011, pp. 27–31.

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, pp. 556–562, 2001.

[10] A. Jansen and P. Niyogi, "A geometric perspective on speech sounds," Tech. Rep., University of Chicago, June 2005.

[11] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[12] A. Cichocki and S.-i. Amari, "Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.

[13] A. Cichocki, S.-i. Amari, R. Zdunek, R. Kompass, G. Hori, and Z. He, "Extended SMART algorithms for non-negative matrix factorization," *Lecture Notes in Artificial Intelligence*, vol. 4029, pp. 548–562, 2006.

[14] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Machine Learning Research 5*, pp. 1457–1469, 2004.

[15] S. Kirkpatrick, C. Gelatt, and M. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, 1983.