UNIVERSAL SPEECH MODELS FOR SPEAKER INDEPENDENT SINGLE CHANNEL SOURCE SEPARATION

*Dennis L. Sun**

Department of Statistics
Stanford University

Gautham J. Mysore

Adobe Research

ABSTRACT

Supervised and semi-supervised source separation algorithms based on non-negative matrix factorization have been shown to be quite effective. However, they require isolated training examples of one or more sources, which is often difficult to obtain. This limits the practical applicability of these algorithms. We examine the problem of efficiently utilizing general training data in the absence of specific training examples. Specifically, we propose a method to learn a universal speech model from a general corpus of speech and show how to use this model to separate speech from other sound sources. This model is used in lieu of a speech model trained on speaker-dependent training examples, and thus circumvents the aforementioned problem. Our experimental results show that our method achieves nearly the same performance as when speaker-dependent training examples are used. Furthermore, we show that our method improves performance when training data of the non-speech source is available.

Index Terms— source separation, non-negative matrix factorization

1. INTRODUCTION

Data-driven approaches [1, 2] have proven quite effective at separating sources in an audio signal. The workhorse behind many methods is non-negative matrix factorization (NMF), which solves the optimization problem:

$$\min_{W,H \geq 0} D(V||WH)$$

where D is a suitably chosen divergence measure, V is the power or magnitude spectrogram, and W, H are the desired factors. Because the factors are constrained to be nonnegative, W and H have natural interpretations as the latent spectral features and the activations of those features in the signal, respectively. A comprehensive survey of this model, including a discussion of different choices of D, is provided in [3].

From this model, a typical pipeline for performing data-driven source separation proceeds as follows [4]. Given isolated training data for the two sources, say speech and noise:

1. Compute the spectrograms V_S and V_N of the speech and noise training data, respectively, as well as the spectrogram V of the mixture signal.
2. Factorize the spectrograms $V_i \approx W_i \tilde{H}_i$, $i = S, N$.
3. Fix the learned spectral features W_S, W_N from above and learn the activations H in the mixture signal:
$$V \approx [W_S \quad W_N] H. \tag{1}$$
4. The activations can then be partitioned into two blocks $H = \begin{bmatrix} H_S \\ H_N \end{bmatrix}$, one corresponding to the speech, the other to the noise. From this, the speech part of the mixture can be recovered as $W_S H_S$. This serves as the estimated speech spectrogram, from which we can obtain the speech waveform estimate by combining it with the mixture phase, and taking the inverse STFT.

The approach described above is known as **supervised** separation. A similar approach is also possible in the (more realistic) scenario where isolated training data is available for only one of the two sources. For example, in speech denoising, it may be possible to obtain isolated noise training data (e.g., when the speaker pauses), but not isolated speech training data. This **semi-supervised** case requires just a slight modification to the above algorithm: instead of learning just H in (1) above, we simultaneously learn W_S and H.

A natural next question is whether the knowledge that the other source is speech can be utilized in some way to improve upon semi-supervised separation or to perform separation when there is no training data of either source. To know a sound class, such as speech, is to have a mathematically useful representation of it. We will learn this representation from data: examples similar to the source we wish to extract.

We refer to such representations as *universal audio models* in analogy to universal background models (UBMs) for speaker verification [5]. Like UBMs, universal audio models also involve pre-training on a large corpus of examples, but models may vary as to what features are learned and how they are used. In the following sections, we propose one model and demonstrate its effectiveness for separating speech and noise.

2. A UNIVERSAL SPEECH MODEL

In this section, we propose a universal speech model based on the principle of block sparsity. We focus on the speech denoising application and discuss a universal speech model, but the same ideas can potentially be applied to any class of sounds.

*This work was performed while interning at Adobe Research.

2.1. Model Training

The block sparsity model decouples the training of the model from its application. In the training stage, a matrix W_i of basis vectors is learned separately for each speaker in the corpus, $i = 1, ..., M$. This can be done using NMF, a probabilistic model, or even by handcrafting the basis vectors, although we used NMF in our experiments. The universal speech model is then obtained by concatenating the learned speech model into a single large matrix:

$$W_S = [W_1 \quad \cdots \quad W_M].$$

To add a speaker to an existing model, we simply learn its basis vectors, independently of how the existing model was learned, thus enabling efficient reuse of data and rendering extensions to the model trivial.

2.2. Model Fitting

If we have a noise model W_N in addition to the universal speech model, separating speech and noise becomes a problem of finding the corresponding activations H_S and H_N as described in Section 1. However, the number of parameters is large, possibly more than the number of observations, so simply finding $H \geq 0$ minimizing $D(V||WH)$ may not yield the best separation results. In high-dimensional settings, appropriate regularization can be an effective strategy to prevent overfitting [6].

Block sparsity refers to one choice of regularization. The intuition is that if the actual speaker or a speaker very similar to the actual speaker is in the universal model, then supervised separation using only the basis vectors for that speaker is close to optimal. This can be achieved by imposing a penalty Ω that induces block sparsity of H_S, where the blocks are the activations of the individual speaker models. The optimization criterion is shown in (2).

$$\min_{W,H \geq 0} D(V||WH) + \lambda \Omega(H_S). \tag{2}$$

For sufficiently large λ, this penalty encourages only one speaker model to be used. At the other extreme, $\lambda = 0$ corresponds to the case described above where the entire universal speech model is used. For λ in between these two extremes, the model is permitted to borrow strength from different models in case a single speaker model is insufficient.

The parameter λ controls the tradeoff between separation and artifacts. For $\lambda = 0$, the reconstructed sources have few artifacts but the separation is poor. As λ increases, separation typically improves at the price of artifacts. See Figure 2 for details. Thus, λ is a tuning parameter with a physical interpretation that an end user could adjust, depending on the requirements of the application.

Block sparsity also provides robustness against poorly fitting speech models by omitting them entirely. Figure 1 depicts the evolution of the activation matrix H as λ increases for a universal speech model consisting of two female speakers and one male speaker ($K = 20$ basis vectors each), as applied to a test mixture of the first female speaker and motorcycle noise ($K = 10$ basis vectors). All coefficients are active for $\lambda = 0$, the male speaker model is dropped first as λ increases, and only the basis vectors of the actual speaker remain for λ sufficiently large. Notice that the noise activations are not penalized and that the group sparsity comes at the price of shrinking the coefficients within each group, a phenomenon that will be discussed in Section 2.3.

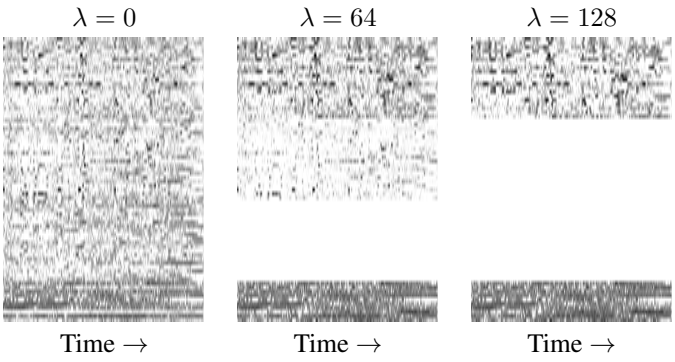

Fig. 1. The activation matrix H for a 3-speaker universal speech model for different values of λ.

Penalty	$\Omega(H_S)$
ℓ_1/ℓ_∞	$\sum_{g=1}^{M} \|\|H_g\|\|_\infty$
ℓ_1/ℓ_2	$\sum_{g=1}^{M} \|\|H_g\|\|_2$
$\log/\ell_1$	$\sum_{g=1}^{M} \log(\epsilon + \|\|H_g\|\|_1)$
ℓ_0/ℓ_1	$\#\{g : \|\|H_g\|\|_1 > 0\}$

Table 1. Examples of common penalties which induce sparsity in the blocks H_g. All norms are elementwise over the matrix entries.

2.3. Related Work and Choice of Ω

Block-sparsity-inducing penalties were introduced in [7] and are also known as *group lasso* or *multitask regression* in the literature. Table 1 describes some common choices for the penalty Ω. In an audio context, [8] used group sparsity as a structural assumption to perform *unsupervised* separation, while [9] examined its application to speaker identification.

The choice of Ω is a delicate issue. Penalties that induce block sparsity typically involve an outer penalty which induces sparsity in the norms of the blocks; by forcing the norms of the blocks to zero, the entries in those blocks are also forced to zero. For example, the ℓ_0/ℓ_2 penalty penalizes the ℓ_0 norm (the number of nonzero components) of the ℓ_2 norms of the blocks. It is ideal in that it penalizes the number of blocks without further shrinking the coefficients, but it is intractable to solve in general. There are a number of relaxations of the ℓ_0 norm which admit tractable solutions, but all involve some shrinkage of the coefficients.

A $\log/\ell_1$ penalty is suggested in [8], for which there are simple multiplicative updates that monotonically decrease the objective. The convex ℓ_1/ℓ_2 penalty is considered in [9], which is fit using heuristic multiplicative updates for which

monotonicity and convergence are still open questions.

2.4. Algorithms

We propose an algorithm for (2) where D is Kullback-Leibler divergence and Ω is $\log/\ell_1$, a problem that will henceforth be referred to as BKL-NMF (block KL-NMF). First, we consider the *supervised* setting, where isolated noise training data is available and hence W is fixed. An iterative algorithm can be derived by majorization-minimization. First, using Jensen's inequality, we can majorize D for any $\sum_k \pi_{ijk} = 1$:

$$D(V||WH) \leq -\sum_{i,j} V_{ij} \sum_k \pi_{ijk} \log W_{ik} H_{kj} + \sum_{i,j} (WH)_{ij} + \text{const.} \quad (3)$$

In particular, we can choose $\pi_{ijk} = \frac{W_{ik}\tilde{H}_{kj}}{\sum_k W_{ik}\tilde{H}_{kj}}$, where $\tilde{H}$ denotes the value of H at the current iteration. Next, since Ω is concave, we can majorize it by its tangent at $\tilde{H}$: $\Omega(H) \leq \Omega(\tilde{H}) + \langle \nabla\Omega(\tilde{H}), H - \tilde{H} \rangle$, which yields

$$\lambda\Omega(H_S) \leq \lambda \sum_g \left\langle \frac{\tilde{H}_g}{\epsilon + ||\tilde{H}_g||_1}, H_g \right\rangle + \text{const.} \quad (4)$$

The majorizing function (3) + (4) can be minimized exactly by setting the gradient equal to 0, leading to efficient multiplicative updates, shown in Algorithm 1. This is an example of a concave-convex procedure (CCCP), for which convergence (albeit not to a global optimum) is known [13].

Algorithm 1 Supervised and Semi-supervised BKL-NMF

inputs V, $W = [W_S \quad W_N]$ (assuming $1^T W = 1$)
initialize H
repeat
  $R \leftarrow V./(WH)$
  $H \leftarrow H.*(W^T R)$
  for $g = 1 : M$ **do**
    $H_g \leftarrow \frac{1}{1 + \lambda/(\epsilon + ||H_g||_1)} H_g$
  end for
  if semi-supervised **then**
    $W_N \leftarrow W_N.*(RH_N^T)$
    $W_N \leftarrow W_N./(11^T W_N)$ (renormalize W)
  end if
until convergence **return** H

$.*$ and $./$ denote componentwise multiplication and division.

Algorithm 1 differs from standard supervised separation algorithms [4] only in the blockwise application of a shrinkage factor. Thus, the universal speech model can be fitted at effectively the same computational cost as standard NMF with KL divergence (hereafter, KL-NMF). In fact, from the end user's perspective, the relevant comparison is between supervised BKL-NMF and semi-supervised KL-NMF—the two options when only noise training examples are available. The latter additionally requires an update of W_N, so as a result, supervised BKL-NMF (for a single λ) can even be faster than semi-supervised KL-NMF.

In the *semi-supervised* setting where no training data of either source is available, only one additional update of the noise model W_N is required. This is reflected in Algorithm 1.

	SDR	K						
	(dB)	5	10	20	30	40	50	100
M	5	9.60	9.85	9.77	9.60	9.49	9.30	8.96
	10	9.82	9.90	9.95	9.64	9.64	9.43	9.02
	20	9.72	9.96	9.92	9.68	9.68	9.58	8.99
	30	9.85	9.84	9.92	9.53	9.66	9.51	8.93
	40	9.92	9.93	9.70	9.54	9.22	9.09	8.50
	50	9.78	10.03	9.78	9.58	9.43	9.19	8.38

Table 2. The optimal (over λ) Signal-to-Distortion Ratio (SDR) for different combinations of number of basis vectors K and number of speakers M in the universal model, averaged over 50 test examples (5 test speakers $\times$ 10 noise examples). The standard error of each estimate was around 0.6.

3. EXPERIMENTS

The experiments described in this section serve two purposes: to determine the optimal parameter settings for the universal speech model and to compare its performance to speaker-dependent models.

3.1. Optimal Parameters for Universal Speech Models

We trained universal speech models with $M = 5$, 10, 20, 30, 40, 50 male speakers chosen randomly from the TIMIT speech corpus using standard KL-NMF, and tested on synthetic mixtures of each of 5 held-out speakers and each of 10 noise examples for a total of 50 test examples. We used the noise examples from [14], which include nonstationary noises such as computer keyboards and birds. The speech and noise signals were normalized to have equal power, i.e., a signal-to-noise ratio of 0 dB.

The other tuning parameter in universal speech models is the number of basis vectors per speaker. For simplicity, we assumed that each speaker model had the same number of basis vectors and considered $K = 5$, 10, 20, 30, 40, 50, 100. For each noise example, we used the optimal number of basis vectors found in [14].

Then, the universal speech model for different values of λ was applied to the mixture signal and the BSS evaluation metrics calculated for each separation [15]. The optimal Signal-to-Distortion Ratio (SDR), a standard single-number summary of separation performance, over the λ was recorded for each K and M. The results are shown in Table 2. Although the variability across examples is quite high (the standard errors of these estimates are around 0.6), the results suggest that the universal speech model is fairly robust with respect to choice of K and M.

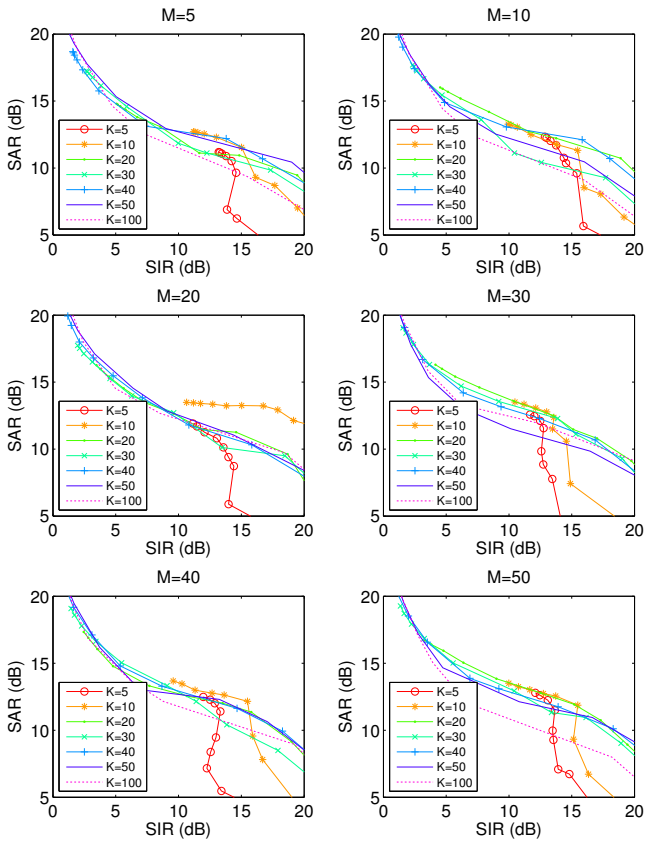

Fig. 2. SIR and SAR tradeoff curves for different settings of number of speakers (M) and number of basis vectors (K), as applied to the mixture of speaker `mthc0` from the TIMIT speech database and motorcycle noise.

The tradeoff between Signal-to-Interference Ratio (SIR) and Signal-to-Artifacts Ratio (SAR) as λ varies, shown in Figure 2 for a particular test mixture, gives a more comprehensive picture of the separation quality. The lines act as Pareto frontiers; one parameter setting dominates another if its tradeoff curve lies up and to the right of the latter's, since then it is possible to simultaneously improve both SIR and SAR. For this particular example, $K = 10$ performs well in general, and while the performance improves as the number of speakers M increases, the gain is modest.

In the experiments that follow, we use a universal model with $K = 10$ and $M = 20$.

3.2. Evaluation of Universal Speech Models

As a comparison, we applied supervised and semi-supervised KL-NMF to each of the 50 test examples from above and recorded the optimal SDR over different choices of the number of speech basis vectors. We applied the universal speech model with the number of basis vectors per speaker fixed at $K = 10$ for each of $M = 20$ speakers and recorded the optimal SDR over different choices of the regularization parameter λ. In both cases, *supervised* indicates that both a speech model (either the universal or the speaker-dependent) and a noise model are used, whereas *semi-supervised* indicates that only a speech model is used.

	KL-NMF	BKL-NMF w/ Univ. Speech Model
supervised	10.23	10.41
semi-supervised	7.22	6.23

Table 3. Comparison of using a speech model trained on speaker dependent training data to using the universal speech model. The SDR is averaged over 50 test examples.

	KL-NMF	BKL-NMF w/ Univ. Speech Model
noise training only	9.27	10.41
no training data	—	6.23

Table 4. Comparison of scenarios based on the type of training data that is provided by the end user. The SDR is averaged over 50 test examples.

As a proof of concept, we first quantified the performance loss of using a universal speech model in place of a speaker-dependent model. The optimal SDRs, averaged over the 50 test examples, are shown in Table 3. We see that the universal speech model provides comparable performance to the speaker-dependent model, even performing slightly better in the supervised case, although this effect is only marginally significant ($t = 2.0$, $p = .05$).

Next, we considered two practical scenarios based on the type of training data that is provided by the end user.

1. Only noise training data is provided: Semi-supervised KL-NMF can be used, where a noise model but no speech model has been provided. This can be compared to supervised separation using BKL-NMF with the universal speech model, which also requires only a noise model from the user. As shown in Table 4, BKL-NMF with the universal speech model performs better ($t = 6.5$, $p = 4 \times 10^{-8}$).

2. No training data is provided: This is a common real-world problem that KL-NMF cannot handle. However, when using BKL-NMF with the universal speech model, we are able to achieve a separation performance of about 6 dB SDR, as shown in Table 4.

4. CONCLUSION

We have proposed a method for performing source separation using general training data in the absence of specific training examples. Our approach learns a model for each example, and uses block sparsity in the fitting to select a subset of models. The resulting speaker-independent model can be fit at about the same computational cost as standard NMF and achieves comparable performance to methods utilizing speaker-dependent training data. Although our exposition has focused on speech denoising, the general idea readily extends to other sources and any number of them.

5. REFERENCES

[1] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *International Conference on Independent Component Analysis and Blind Signal Separation*, 2004.

[2] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, 2007.

[3] C. Févotte, N. Bertin, and J.L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, 2009.

[4] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *International Conference on Independent Component Analysis and Signal Separation*, 2007.

[5] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, 2000.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2nd edition, 2009.

[7] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, no. 1, 2005.

[8] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[9] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *Interspeech*, 2012.

[10] E.J. Candès, M.B. Wakin, and S.P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[11] A. Lefèvre, *Dictionary learning methods for single-channel source separation*, Ph.D. thesis, Ecole Normale Supérieure de Cachan, 2012.

[12] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, 2011.

[13] B. Sriperumbudur and G. Lanckriet, "On the convergence of the concave-convex procedure," in *Advances in Neural Information Processing Systems*, 2009.

[14] Z. Duan, G. J. Mysore, and P. Smaragdis, "Online plca for real-time semi-supervised source separation," in *International Conference on Latent Variable Analysis and Source Separation (LVA/ICA)*. 2012, Springer.

[15] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2006.