Single-Channel Speech Separation using Sparse Non-Negative Matrix Factorization

Mikkel N. Schmidt and Rasmus K. Olsson

Informatics and Mathematical Modelling
Technical University of Denmark
2800 Lyngby, Denmark
mns,rko@imm.dtu.dk

Abstract

We apply machine learning techniques to the problem of separating multiple speech sources from a single microphone recording. The method of choice is a sparse non-negative matrix factorization algorithm, which in an unsupervised manner can learn sparse representations of the data. This is applied to the learning of personalized dictionaries from a speech corpus, which in turn are used to separate the audio stream into its components. We show that computational savings can be achieved by segmenting the training data on a phoneme level. To split the data, a conventional speech recognizer is used. The performance of the unsupervised and supervised adaptation schemes result in significant improvements in terms of the target-to-masker ratio.

1. Introduction

A general problem in many applications is that of extracting the underlying sources from a mixture. A classical example is the so-called cocktail-party problem in which the problem is to recognize or isolate what is being said by an individual speaker in a mixture of speech from various speakers. A particular difficult version of the cocktail-party problem occurs when only a single-channel recording is available, yet the human auditory system solves this problem for us. Despite its obvious possible applications in, e.g., hearing aids or as a preprocessor to a speech recognition system, no machine has been built, which solves this problem in general.

Within the signal processing and machine learning communities, the single channel separation problem has been studied extensively, and different parametric and non-parametric signal models have been proposed.

Hidden Markov models (HMM) are quite powerful for modelling a single speaker. It has been suggested by Roweis [1] to use a factorial HMM to separate mixed speech. Another suggestion by Roweis is to use a factorial-max vector quantizer [2]. Jang and Lee [3] use independent component analysis (ICA) to learn a dictionary for sparse encoding [4], which optimizes an independence measure across the encoding of the different sources. Pearlmutter and Olsson [5] generalize these results to overcomplete dictionaries, where the number of dictionary elements is allowed to exceed the dimensionality of the data. Other methods learn spectral dictionaries based on different types of non-negative matrix factorization (NMF) [6]. One idea is to assume a convolutive sum mixture, allowing the basis functions to capture time-frequency structures [7, 8].

A number researchers have taken ideas from the computational auditory scene analysis (CASA) literature, trying to incorporate various grouping cues of the human auditory system in speech separation algorithms [9, 10]. In the work by Ellis and Weiss [11] careful consideration is given to the representation of the audio signals so that the perceived quality of the separation is maximized.

In this work we propose to use the sparse non-negative matrix factorization (SNMF) [12] as a computationally attractive approach to sparse encoding separation. As a first step, overcomplete dictionaries are estimated for different speakers to give sparse representations of the signals. Separation of the source signals is achieved by merging the dictionaries pertaining to the sources in the mixture and then computing the sparse decomposition. We explore the significance of the degree of sparseness and the number of dictionary elements. We then compare the basic unsupervised SNMF with a supervised application of the same algorithm in which the training data is split into phoneme-level subproblems, leading to considerable computational savings.

The article is organized as follows: First, the separation method based on SNMF is explained in details, and we elaborate on the idea of computing the SNMF on individual phonemes. This is followed by simulations demonstrating the usefulness of the algorithm on a speech separation task. We conclude with a brief discussion and suggest future improvements of the approach.

2. Method

In the following, we consider modelling a magnitude spectrogram representation of a mixed speech signal. We represent the speech signal in the non-negative Mel spectrum magnitude domain, as suggested by Ellis and Weiss [11].

Here we posit that the spectrogram can be sparsely represented in an overcomplete basis,

$$\mathbf{Y} = \mathbf{D}\mathbf{H} \tag{1}$$

that is, each data point held in the columns of $\mathbf{Y}$ is a linear combination of few columns of $\mathbf{D}$. The dictionary, $\mathbf{D}$, can hold arbitrarily many columns, and the code matrix, $\mathbf{H}$, is sparse. Furthermore, we assume that the mixture signal is a sum of R source signals

$$\mathbf{Y} = \sum_{i}^{R} \mathbf{Y}_i.$$

The basis of the mixture signal is then the concatenation of the source dictionaries, $\mathbf{D} = [\mathbf{D}_1 \ldots \mathbf{D}_i \ldots \mathbf{D}_R]$, and the complete code matrix is the concatenation of the source-individual codes,

$\mathbf{H} = \left[\mathbf{H}_1^\top \ldots \mathbf{H}_i^\top \ldots \mathbf{H}_R^\top\right]^\top$. By enforcing the sparsity of the code matrix, $\mathbf{H}$, it is possible to separate $\mathbf{Y}$ into its sources if the dictionaries are diverse enough.

As a consequence of the above, two connected tasks have to be solved: 1) the learning of source-specific dictionaries that yield sparse codes, and, 2) the computing of sparse decompositions for separation. We will use the sparse non-negative matrix factorization method proposed by Eggert and Körner [12] for both tasks.

2.1. Sparse Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) computes the decomposition in Equation (1) subject to the constraints that all matrices are non-negative, leading to solutions that are parts-based or sparse [6]. However, the basic NMF does not provide a well-defined solution in the case of overcomplete dictionaries, when the non-negativity constraints are not sufficient to obtain a sparse solution. The sparse non-negative matrix factorization (SNMF) optimizes the cost function

$$E \quad = \quad ||\mathbf{Y} - \bar{\mathbf{D}}\mathbf{H}||_F^2 + \lambda \sum_{ij} \mathbf{H}_{ij} \quad \text{s.t.} \quad \mathbf{D}, \mathbf{H} \geq \mathbf{0} \qquad (2)$$

where $\bar{\mathbf{D}}$ is the column-wise normalized dictionary matrix. This cost function is the basic NMF quadratic cost augmented by an L_1 norm penalty term on the coefficients in the code matrix. The parameter, λ, controls the degree of sparsity. Any algorithm that optimizes Equation (2) can be regarded as computing a maximum posterior (MAP) estimate given a Gaussian likelihood function and a one-sided Laplacian prior distribution over $\mathbf{H}$. The SNMF can be computed by alternating updates of $\mathbf{D}$ and $\mathbf{H}$ by the following rules [12]

$$\begin{aligned} \mathbf{H}_{ij} \quad &\leftarrow \quad \mathbf{H}_{ij} \bullet \frac{\mathbf{Y}_i^\top \bar{\mathbf{D}}_j}{\mathbf{R}_i^\top \bar{\mathbf{D}}_j + \lambda} \\ \mathbf{D}_j \quad &\leftarrow \quad \mathbf{D}_j \bullet \frac{\sum_i \mathbf{H}_{ij} \left[\mathbf{Y}_i + (\mathbf{R}_i^\top \bar{\mathbf{D}}_j)\bar{\mathbf{D}}_j\right]}{\sum_i \mathbf{H}_{ij} \left[\mathbf{R}_i + (\mathbf{V}_i^\top \bar{\mathbf{D}}_j)\bar{\mathbf{D}}_j\right]} \end{aligned}$$

where $\mathbf{R} = \mathbf{DH}$, and the bold operators indicate pointwise multiplication and division.

We first apply SNMF to learn dictionaries of individual speakers. To separate speech mixtures we keep the dictionary fixed and update only the code matrix, $\mathbf{H}$. The speech is then separated by computing the reconstruction of the parts of the sparse decomposition pertaining to each of the used dictionaries. In cases, when the identities of the speakers within a given mixture are unknown, they can be estimated as the combination of dictionaries that minimize Equation (2).

2.2. Two Ways to Learn Sparse Dictionaries

We study two approaches to learning sparse dictionaries, see Figure 1. The first is a direct, unsupervised approach where the dictionary is learned by computing the SNMF on a large training data set of a single speaker. The second approach is to first segment the training data according to phoneme labels obtained by speech recognition software based on a hidden Markov model. Then, a sparse dictionary is learned for each phoneme and the final dictionary is constructed by concatenating the individual phoneme dictionaries. As a consequence, a smaller learning problem is addressed by the SNMF for each of the phonemes.

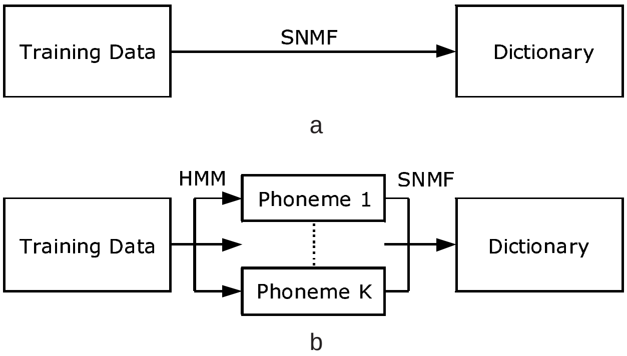

Figure 1: Two approaches for learning sparse dictionaries of speech. The first approach (a) is to learn the dictionary from a sparse non-negative matrix factorization of the complete training data. The second approach (b) is to segment the training data into individual phonemes, learn a sparse dictionary for each phoneme, and compute the dictionary by concatenating the individual phoneme dictionaries.

The computational savings associated with this divide-and-conquer approach are significant. Since the running time of the SNMF scales with the size of the training data and the number of elements in the dictionary, dividing the problem into SNMF subproblems for each phoneme reduces the overall computational burden by a factor corresponding to the number of phonemes. For example, if the data is split into 40 phonemes, we need to solve 40 SNMF subproblems each with a complexity of $1/40^2$ compared to the full SNMF problem. In addition to this, since the phoneme SNMF subproblems are much smaller than the total SNMF problem, a faster convergence of the iterative SNMF algorithm can be expected. These advantages makes it desirable to compare the quality of sparse dictionaries estimated by the two methods.

3. Simulations

Part of the Grid Corpus [13] was used for evaluating the proposed method for speech separation. The Grid Corpus consists of simple structured sentences from a small vocabulary, and has 34 speakers and 1000 sentences per speaker. Each utterance is a few seconds and word level transcriptions are available. We used half of the corpus as a training set.

3.1. Phoneme Transcription

First, we used speech recognition software to generate phoneme transcriptions of the sentences. For each speaker in the corpus a phoneme-based hidden Markov model (HMM) was trained using the HTK toolkit[1]. The HMM's were used to compute an alignment of the phonemes in each sentence, taking the pronunciations of each word from the British English Example Pronunciation (BEEP) dictionary[2]. This procedure provided phoneme-level transcriptions of each sentence. In order to evaluate the quality of the phoneme alignment, the automatic phoneme transcription was compared to a manual transcription for a few sentences. We found that the automatic phoneme alignment in general was quite

[1] Avaiable from htk.eng.cam.ac.uk.

[2] Available by anonymous ftp from svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz.

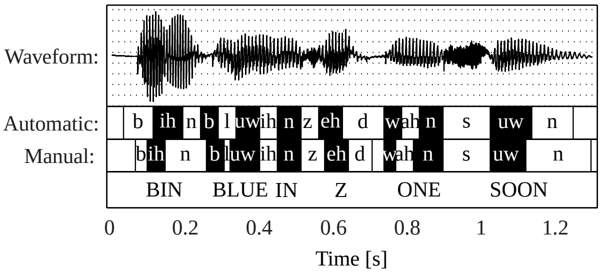

Figure 2: The automatic phoneme transcription as computed by the trained hidden Markov model (HMM) for an example sentence from the Grid Corpus. A manual transcription is provided for comparison, confirming the conventional hypothesis that the HMM is a useful tool in segmenting a speech signal into its phonemes.

reasonable. An example is given in Figure 2.

3.2. Preprocessing and Learning Dictionaries

We preprocessed the speech data in a similar fashion to Ellis and Weiss [11]: the speech was prefiltered with a high-pass filter, $1-0.95z^{-1}$, and the STFT was computed with an analysis window of $32\mathrm{ms}$, corresponding to 800 samples at a sample rate of $25\mathrm{kHz}$. An overlap of 50 percent was used between frames. This yielded a spectrogram with 401 frequency bins which was then mapped into 80 frequency bins on the Mel scale. The training set was reweighted so that all frames containing energy above a threshold were normalized by their standard deviation. The resulting magnitude Mel-scale spectrogram representation was employed in the experiments.

In order to assess the effects of the model hyper-parameters and the effect of splitting the training data according the phoneme transcriptions, a subset of four male and four female speakers were extracted from the Grid Corpus. We constructed a set of 64 mixed sentences by mixing two randomly selected sentences for all combinations of the eight selected test speakers.

Two different sets of dictionaries were estimated for each speaker. The first set was computed by concatenating the spectrograms for each speaker and computing the SNMF on the complete training data for that speaker. The second set was computed by concatenating the parts of the training data corresponding to each phoneme for each speaker, computing the SNMF for each phoneme spectrogram individually, and finally concatenating the individual phoneme dictionaries. To save computation, only 10 percent of the training set was used to train the dictionaries. In a Matlab environment running on a $1.6\mathrm{GHz}$ Intel processor the computation of the SNMF for each speaker took approximately 30 minutes, whereas the SNMFs for individual phonemes were computed in a few seconds. The algorithm was allowed to run for maximally 500 iterations or until convergence as defined by the relative change in the cost function. Figure 3 shows samples from a dictionary which was learned using SNMF on the phoneme-segmented training data for a female speaker. The dictionaries were estimated for four different levels of sparsity, $\lambda = \{0.0001, 0.001, 0.01, 0.1\}$, and four different dictionary sizes, $N = \{70, 140, 280, 560\}$. This was done for both the complete and the phoneme-segmented training data.

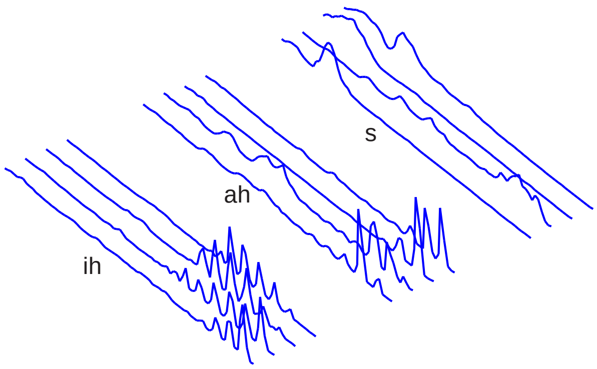

Figure 3: A few samples of columns of phoneme dictionaries learned from female speech. The SNMF was applied to data, which had been phoneme-labelled by a speech recognizer. Not surprisingly, the basis functions exhibit the some general properties of the respective phonemes, and additional variation is captured by the algorithm, such as the fundamental frequency in the case of voiced phonemes.

	Complete	Segmented
Same gender	4.8±0.4 dB	4.3±0.3 dB
Opp. gender	6.6±0.3 dB	6.4±0.3 dB

Table 1: Average signal-to-noise ratio (SNR) of the separated signals for dictionaries trained on the complete speech spectrograms and on individual phonemes. Dictionaries were learned with $N = 560$ and $\lambda = 0.1$.

3.3. Speech Separation

For each test sentence, we concatenated the dictionaries of the two speakers in the mixture, and computed the code matrix using the SNMF updates. Then, we reconstructed the individual magnitude spectra of the two speakers and mapped them from the Mel-frequency domain into the linear frequency STFT domain. Separated waveforms were computed by spectral masking and spectrogram inversion, using the original phase of the mixed signal. The separated waveforms were then compared with the original clean signals, computing the signal-to-noise ratio.

The results in Figure 4 show that the quality of separation increases with N. This agrees well with the findings of Ellis and Weiss [11]. Furthermore, the choice of sparsity, λ, is important for the performance of the separation method, especially in the case of unsegmented data. The individual phoneme-level dictionaries are so small in terms of N that the gain from enforcing sparsity in the NMF is not as significant; the segmentation in itself sparsifies the dictionary to some extend. Table 1 shows that the method works best for separating speakers of opposite gender, as would be expected. Audio examples are available at mikkelschmidt.dk .

3.4. Speaker Identification

We further studied how the sparse dictionaries can be used to identify the speaker. We mapped each mixed sentence onto each combination of dictionaries by concatenating the two dictionaries and computing the SNMF only updating the code matrix. The quality of fit between the mixed signal and the combined dictionary is

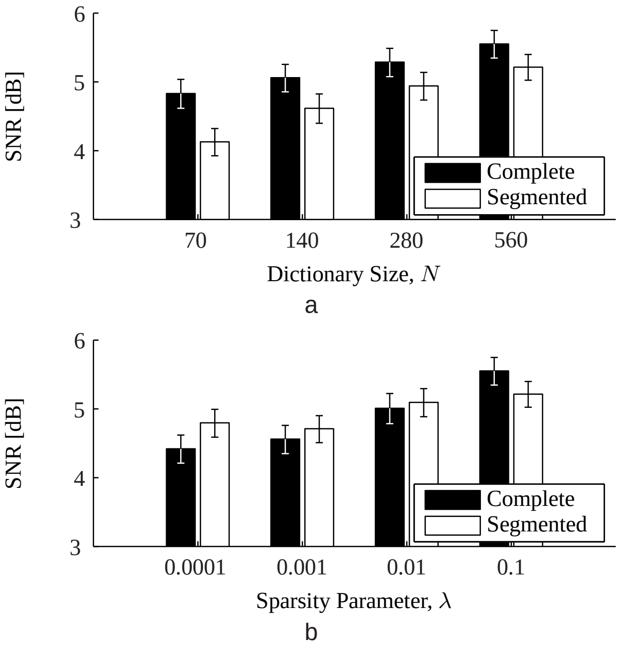

Figure 4: Average signal-to-noise ratio (SNR) of the separated signals for dictionaries trained on the complete speech spectrograms and on individual phonemes, (a) as a function of the dictionary size, N, with sparsity $\lambda = 0.1$, and (b) as a function of the sparsity with $N = 560$.

reflected in the final value of the SNMF cost function. The combination of dictionaries which gave the lowest cost is a good estimate of the identity of the two speakers in the mixture. In our simulations, this estimate was correct 95 percent of the time.

4. Discussion and Outlook

In this work, we have successfully applied sparse non-negative matrix factorization (SNMF) to the problem of monaural speech separation and speaker identification.

The SNMF learns large overcomplete dictionaries in an unsupervised fashion, leading to a more sparse representations of individual speakers than for example the basic NMF. Inspection of the dictionaries reveals that they capture fundamental properties of speech, in fact they learn basis functions that resemble phonemes. This has lead us to adopt a working hypothesis that the learning of signal dictionaries on a phoneme level is a computational shortcut to the goal, leading to similar performance. Our experiments show, that the practical performance of sparse dictionaries learned in this way performs only slightly worse than dictionaries learned on the complete dataset. In future work, we hope to benefit further from the phoneme labelling of the dictionaries in formulating transitional models in the encoding space of the SNMF, hopefully matching the dynamics of speech.

Our results confirm that it is viable to learn personalized dictionaries and apply them blindly, that is, when the identities of the speakers are unknown. We are currently investigating methods to more efficiently determine the active sources in a mixture, rather than exhaustively evaluating all possibilities.

An issue that we are currently studying is that of applying the proposed single-channel speech separator to the task of speech recognition. A major obstacle in this connection is to overcome the generally high sensitivity of speech recognizers to noise and, in particular, the artifacts created by signal enhancement algorithms. A possible answer to this challenge is to train the speech recognizer on data that contains these artifacts, more specifically on "separated" speech sources.

5. Acknowledgements

This work was made possible in part by funding from Oticon Fonden. We would like to thank Lars Kai Hansen and Jan Larsen for fruitful discussions. We gratefully acknowledge Dan Ellis for making useful software available at his homepage, www.ee.columbia.edu/ dpwe/ .

6. References

[1] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems*, 2001, pp. 793–799.

[2] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Eurospeech*, 2003, pp. 1009–1012.

[3] G. J. Jang and T. W. Lee, "A maximum likelihood approach to single channel source separation," *Journal of Machine Learning Research*, vol. 4, pp. 1365–1392, 2003.

[4] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[5] B. A. Pearlmutter and R. K. Olsson, "Algorithmic differentiation of linear programs for single-channel source separation," in *Machine Learning and Signal Processing, in submission*, 2006.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[7] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," in *Statistical and Perceptual Audio Processing (SAPA)*, 2004.

[8] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *ICA*, 2005.

[9] B. A. Pearlmutter and A. M. Zador, "Monaural source separation using spectral cues," in *ICA*, 2004, pp. 478–485.

[10] F. Bach and M. I. Jordan, "Blind one-microphone speech separation: A spectral learning approach," in *Advances in Neural Information Processing Systems 17*, 2005, pp. 65–72.

[11] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *ICASSP*, 2006.

[12] J. Eggert and E. Körner, "Sparse coding and nmf," in *Neural Networks*. 2004, vol. 4, pp. 2529–2533, IEEE.

[13] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *submitted to JASA*.