Discovering Convolutive Speech Phones using Sparseness and Non-Negativity Constraints

Paul D. O'Grady and Barak A. Pearlmutter

Hamilton Institute,
National University of Ireland Maynooth,
Co. Kildare,
Ireland.
paul.ogrady@nuim.ie barak@cs.nuim.ie
http://www.hamilton.ie/paul

Abstract. Discovering a representation that allows auditory data to be parsimoniously represented is useful for many machine learning and signal processing tasks. Such a representation can be constructed by Non-negative Matrix Factorisation (NMF), which is a method for finding parts-based representations of non-negative data. Here, we present an extension to convolutive NMF that includes a sparseness constraint. In combination with a spectral magnitude transform of speech, this method extracts speech phones (and their associated sparse activation patterns), which we use in a supervised separation scheme for monophonic mixtures.

1 Introduction

A preliminary step in many data analysis tasks is to find a suitable representation of the data. Typically, methods exploit the latent structure in the data. For example, ICA reduces the redundancy of the data by projecting the data onto its independent components, which can be discovered by maximising a statistical measure such as independence or non-Gaussianity.

Non-negative matrix factorisation (NMF) approximately decomposes a non-negative matrix $\mathbf{V}$ into a product of two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ [1, 2]. NMF is a parts-based approach that does not make a statistical assumption about the data. Instead, it assumes that for the domain at hand, negative numbers would be physically meaningless. Data that contains negative components, for example audio, must be transformed into a non-negative form before NMF can be applied. Here, we use a magnitude spectrogram. Spectrograms have been used in audio analysis for many years and in combination with NMF have been applied to a variety of problems such as sound separation [3] and automatic transcription of music [4].

In this paper, we combine a previous convolutive extension of NMF [3], which identifies auditory objects with time-varying spectra, with a sparseness constraint, and apply the resulting algorithm to the analysis of speech. The paper is structured as follows: We overview of convolutive NMF in Section 2 and present sparse convolutive NMF in Section 3. In Section 4 we apply sparse convolutive

NMF to speech spectra, and extract phones that have sparse activation patterns. We use these phones in a supervised separation scheme for monophonic mixtures, and demonstrate the superior separation performance achieved over those extracted by convolutive NMF in Section 5.

2 Convolutive NMF

NMF [2] is a linear non-negative approximate factorisation, and is formulated as follows. Given a non-negative $M \times N$ matrix $\mathbf{V} \in \mathbb{R}^{\geq 0, M \times N}$ the goal is to approximate $\mathbf{V}$ as a product of two non-negative matrices $\mathbf{W} \in \mathbb{R}^{\geq 0, M \times R}$ (basis) and $\mathbf{H} \in \mathbb{R}^{\geq 0, R \times N}$ (activations), $\mathbf{V} \approx \mathbf{WH}$, where $R \leq M$, such that the reconstruction error is minimised. For our purposes we require a convolutive basis, such a model has previously been used to extend NMF [3], which we review in this section.

In conventional NMF each object is described by its spectrum and corresponding activation in time, while for convolutive NMF each object has a sequence of successive spectra and corresponding activation pattern across time. The conventional NMF model is extended to the convolutive case:

$$\mathbf{V} \approx \sum_{t=0}^{T_o-1} \mathbf{W}_t \overset{t\rightarrow}{\mathbf{H}}, \qquad v_{ik} \approx \sum_{t=0}^{T_o-1} \sum_{j=1}^{R} w_{ijt} \overset{t\rightarrow}{(h_{jk})}, \tag{1}$$

where T_o is the length of each spectrum sequence and the j-th column of $\mathbf{W}_t$ describes the spectrum of the j-th object t time steps after the object has begun. The function $\overset{i\rightarrow}{(\cdot)}$ denotes a column shift operator that moves its argument i places to the right; as each column is shifted off to the right the leftmost columns are zero filled. Conversely, the $\overset{\leftarrow i}{(\cdot)}$ operator shifts columns off to the left, with zero filling on the right. We use the beta divergence, which is a parameterisable divergence, as the reconstruction objective,

$$D_{\mathrm{BD}}(\mathbf{V} \| \mathbf{\Lambda}, \beta) = \sum_{ik} \Big(v_{ik} \frac{v_{ik}^{\beta-1} - [\mathbf{\Lambda}]_{ik}{}^{\beta-1}}{\beta(\beta-1)} + [\mathbf{\Lambda}]_{ik}{}^{\beta-1} \frac{[\mathbf{\Lambda}]_{ik} - v_{ik}}{\beta} \Big), \tag{2}$$

where β controls reconstruction penalty and $\mathbf{\Lambda}$ is the current estimate of $\mathbf{V}$, $\mathbf{\Lambda} = \sum_{t=0}^{T_o-1} \mathbf{W}_t \overset{t\rightarrow}{\mathbf{H}}$. The choice of the β parameter depends on the statistical distribution of the data, and requires prior knowledge. For $\beta = 2$, Squared Euclidean Distance is obtained; for $\beta \to 1$, the divergence tends to the Kullback-Leibler Divergence; and for $\beta \to 0$, it tends to Itakura-Saito Divergence. It is evident that Eq. 1 can be viewed as a summation of T_o conventional NMF operations. Consequently, as opposed to updating two matrices ($\mathbf{W}$ and $\mathbf{H}$) as in conventional NMF, $T_o + 1$ matrices require an update ($\mathbf{W}_0, \ldots, \mathbf{W}_{T_o-1}$ and $\mathbf{H}$). The resultant convolutive NMF update equations are

$$w_{ijt} \leftarrow w_{ijt} \frac{\sum_{k=1}^{T} (v_{ik} / [\mathbf{\Lambda}]_{ik}^{2-\beta}) \overset{t\rightarrow}{h}_{jk}}{\sum_{k=1}^{T} [\mathbf{\Lambda}]_{ik}^{\beta-1} \overset{t\rightarrow}{h}_{jk}}, \qquad h_{jk} \leftarrow h_{jk} \frac{\sum_{i=1}^{M} w_{ijt} (v_{ik} / [\mathbf{\Lambda}]_{ik}^{2-\beta})}{\sum_{i=1}^{M} w_{ijt} [\overset{\leftarrow t}{\mathbf{\Lambda}}]_{ik}^{\beta-1}}. \tag{3}$$

where $\mathbf{H}$ is updated to the average result of its updates for all t. When $T = 1$ this reduces to conventional NMF.

3 Sparse Convolutive NMF

Combining our reconstruction objective (Eq. 2) with a sparseness constraint on $\mathbf{H}$ results in the following objective function:

$$G(\mathbf{V}\|\mathbf{\Lambda}, \mathbf{H}, \beta) = D_{\mathrm{BD}}(\mathbf{V}\|\mathbf{\Lambda}, \beta) + \lambda \sum_{jk} h_{jk}, \tag{4}$$

where the left term of the objective function corresponds to convolutive NMF, and the right term is an additional constraint on $\mathbf{H}$ that enforces sparsity by minimising the L_1-norm of its elements. The parameter λ controls the trade off between sparseness and accurate reconstruction.

3.1 Basis Normalisation

The objective of Eq. 4 creates a new problem: The right term is a strictly increasing function of the absolute value of its argument, so it is possible that the objective can be decreased by scaling $\mathbf{W}_t$ up and $\mathbf{H}$ down ($\mathbf{W}_t \mapsto \alpha\mathbf{W}_t$ and $\mathbf{H} \mapsto (1/\alpha)\mathbf{H}$, with $\alpha > 1$). This situation does not alter the left term in the objective function, but will cause the right term to decrease, resulting in the elements of $\mathbf{W}_t$ growing without bound and $\mathbf{H}$ tending toward zero. Consequently, the solution arrived at by the optimisation algorithm is not influenced by the sparseness constraint.

To avoid the scaling misbehaviour of Eq. 4 another constraint is needed; by normalising the convolutive bases we can control the scale of the elements in $\mathbf{W}_t$ and $\mathbf{H}$. Normalisation is performed for each object matrix, $\mathbf{W}_j$, by rescaling it to the unit L_2-norm, $\bar{\mathbf{W}}_j = \frac{\mathbf{W}_j}{\|\mathbf{W}_j\|}, j = 1, \ldots, R$, where the matrix $\mathbf{W}_j$ is constructed from the j-th column of $\bar{\mathbf{W}}_t$ at each time step, $t = 0, 1, \ldots, T_o - 1$.

3.2 Multiplicative Updates

Multiplicative updates can be obtained by including the normalisation requirement in the objective. Previously, this has been achieved for conventional NMF using the Squared Euclidean Distance reconstruction objective [5]. Here, we present the multiplicative updates for a convolutive NMF algorithm utilising beta divergence. Our new reconstruction objective is a modification of Eq. 2 where each object, $\mathbf{W}_j$, is normalised, $\bar{\mathbf{W}}_j$, resulting in the following generative model: $\mathbf{\Delta} = \sum_{t=0}^{T_o-1} \sum_{j=1}^{R} \bar{\mathbf{w}}_{jt} (\overrightarrow{\mathbf{h}_j}^{t})$. By substituting $\mathbf{\Lambda}$ for $\mathbf{\Delta}$ in Eq. 4 we obtain the following multiplicative update rule for $\mathbf{H}$,

$$h_{jk} \leftarrow h_{jk} \frac{\sum_{i=1}^{M} \bar{w}_{ijt} (v_{ik} / [\overleftarrow{\mathbf{\Delta}}^{t}]_{ik}^{2-\beta})}{\sum_{i=1}^{M} \bar{w}_{ijt} [\overleftarrow{\mathbf{\Delta}}^{t}]_{ik}^{\beta-1} + \lambda}, \tag{5}$$

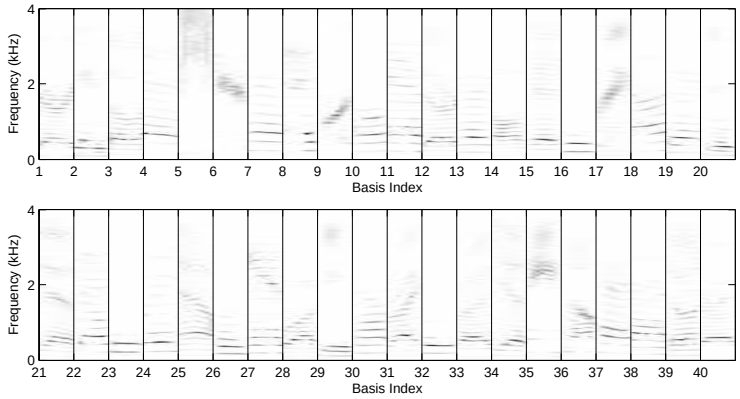

Fig. 1. A collection of 40 phone-like basis functions for a mixture of a male (`DMT0`) and female speaker (`SMA0`) taken from the TIMIT speech database.

and update for $\mathbf{W}$,

$$w_{ijt} \leftarrow w_{ijt} \frac{\sum_{k=1}^{T} \overset{t\rightarrow}{h}_{jk} \left[(v_{ik} / [\mathbf{\Delta}]_{ik}^{2-\beta}) + \bar{w}_{ijt} (\bar{w}_{ijt} [\mathbf{\Delta}]_{ik}^{\beta-1}) \right]}{\sum_{k=1}^{T} \overset{t\rightarrow}{h}_{jk} \left[[\mathbf{\Delta}]_{ik}^{\beta-1} + \bar{w}_{ijt} (\bar{w}_{ijt} (v_{ik} / [\mathbf{\Delta}]_{ik}^{2-\beta})) \right]}. \tag{6}$$

4 Sparse Convolutive NMF on Speech Spectra

We apply sparse convolutive NMF to speech, and present a learned basis for the sparse representation of speech using the TIMIT database. Recently, such work has been presented for convolutive NMF [6].

4.1 Discovering a Phone-like Basis

To illustrate the differences between the phones extracted by convolutive NMF and sparse convolutive NMF we perform the following experiment for both algorithms: We take around 15 seconds of speech from a male speaker (`DMT0`) and female speaker (`SMA0`) to create a contiguous mixture. The data is normalised to unit variance, down-sampled from 16 kHz to 8 kHz and a magnitude spectrogram of the data is constructed. We use a FFT frame size of 512, a frame overlap of 384 and a hamming window to reduce the presence of sidelobes. We extract 40 bases, $R = 40$, with a temporal extent of 0.176 seconds, $T_o = 8$, and run convolutive NMF (with $\beta = 1$) for 200 iterations. The extracted bases are presented in Figure 1. The experiment is repeated for sparse convolutive NMF with $\lambda = 15$, and the corresponding bases are presented in Figure 2.

For convolutive NMF, it is evident that the extracted bases correspond to speech phones. The verification of which, can be achieved by listening to an audible reconstruction. Most of the phones represent harmonic series with differing

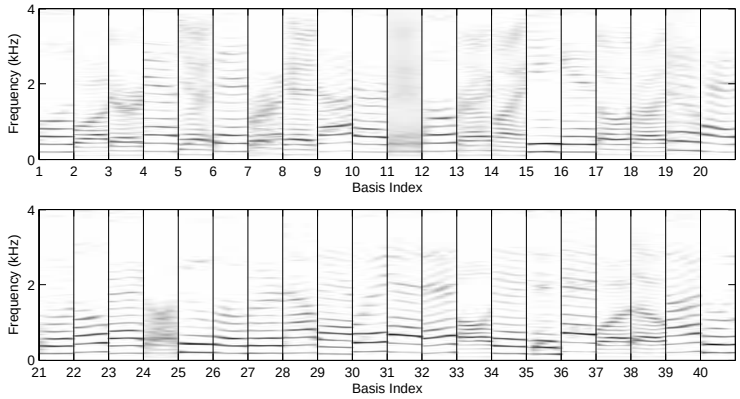

Fig. 2. A collection of 40 phone-like basis functions for a a mixture of a male (`DMT0`) and female speaker (`SMA0`) taken from the TIMIT speech database. The basis is extracted using Spare Convolutive NMF with $\lambda = 15$.

pitch inflections, while a smaller subset of phones contain wideband components that correspond to consonant sounds. It is evident for the harmonic phones that some bases have harmonics that are spaced much closer together, which is indicative of a lower pitched male voice, while others are farther apart, indicating a higher pitched female voice. Therefore, it is evident that the extracted phones correspond to either the male or female speaker, which indicates that the timbral characteristics of the male and female speaker are sufficiently different, such that phones that are representative of both cannot be extracted. Although, this may not be true for the consonant phones.

By placing a sparseness constraint on the activations of the basis functions, we specify that the expressive power of each basis be extended such that it is capable of representing phones parsimoniously, much like an over-complete dictionary. The result is that the extracted phones exhibit a structure that is rich in phonetic content, where harmonics at higher frequencies have a much greater intensity than seen in the phones extracted by convolutive NMF. This reflects the requirement that the basis functions in our new sparse phone set, must contain enough features to produce a parsimonious activation pattern. Analysis of the male and female sparse phone set reveals another important difference between the two speakers. In addition to difference in harmonic spacing, it is evident that the structure of the male phones are of a more complex nature, where changes over time are much more varied than for the female phone set.

5 Supervised Method for the Separation of Speakers

As illustrated in our previous experiments, the structure of the bases that are extracted from the speech data are uniquely dependent on the speaker (given the same algorithm parameters). In the context of speech separation, it is not

unreasonable to expect that the bases extracted for a specific speaker adequately characterise the speaker, such that they can be used to discriminate them from other speakers. For a monophonic mixture where a number speakers are added together, it is possible to separate the speakers in the mixture by constructing an individual magnitude spectrogram for each speaker, using the phones specific to that speaker. Specifically, we use the following procedure for the separation of a known male and female speaker from a monophonic mixture:

1. Obtain training data for the male, $s_m(t)$, and female, $s_f(t)$, speaker, create a magnitude spectrogram for both, and extract corresponding phone sets, $\mathbf{W}_t^m$ and $\mathbf{W}_t^f$, using sparse convolutive NMF.
2. Construct a combined basis set $\mathbf{W}_t^{mf} = [\mathbf{W}_t^m|\mathbf{W}_t^f]$, which results in a basis that is twice as big as R.
3. Take a mixture that is composed of two unknown sentences voiced by our selected speakers, and create a magnitude spectrogram of the mixture. Fit the mixture to $\mathbf{W}_t^{mf}$ by performing sparse convolutive NMF with $\mathbf{W}_t$ fixed to $\mathbf{W}_t^{mf}$, and learn only the associated activations $\mathbf{H}$.
4. Partition $\mathbf{H}$ such that the activations are split into male, $\mathbf{H}^m$, and female, $\mathbf{H}^f$, parts that correspond to their associated bases, $\mathbf{H} = [\mathbf{H}^m|\mathbf{H}^f]$.
5. Construct a magnitude spectrogram for both speakers, using their respective bases and activations: $\mathbf{S}^m = \sum_{t=o}^{T_o-1} \mathbf{W}^m\mathbf{H}^m$ and $\mathbf{S}^f = \sum_{t=o}^{T_o-1} \mathbf{W}^f\mathbf{H}^f$.
6. Use the phase information from the mixture to create an audible reconstruction for both speakers.

This procedure may also be used for convolutive NMF, and can be generalised for more than two speakers, and speakers of the same gender.

5.1 Separation Experiments

Here, we compare the separation performance of convolutive NMF and sparse convolutive NMF. For an extensive study of the relationship between parameter selection and separation performance for convolutive NMF, see [6].

We select three male (`ABC0`, `BJV0`, `DWM0`) and three female (`EXM0`, `KLH0`, `REH0`) speakers from the TIMIT database, and create a training set for each that includes all but one sentence voiced by that speaker. We artificially generate a monophonic mixture by summing the remaining sentences for a selected male female pair, generating a total of nine mixtures in this way. More formally, each sentence pair is normalised to unit variance, down-sampled from 16 kHz to 8 kHz, and summed together. A magnitude spectrogram of each mixture is constructed using a FFT frame size of 512, a frame overlap of 256 and a hamming window.

The separation performance for both algorithms is evaluated for each mixture over a selection of values for R ($R = [40\ 80\ 140\ 220]$). For both algorithms the temporal extent of each phone is set to 0.224 seconds ($T_o = 6$), the number of iterations is 150, β is set to 1 and each experiment is repeated for 10 Monte Carlo runs. For convolutive NMF, a total of 24 speaker phone sets are extracted

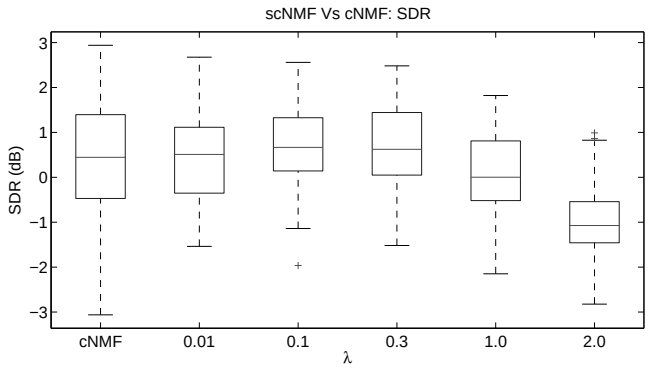

Fig. 3. A comparison of the SDR results obtained by convolutive and sparse convolutive NMF: Box plots are used to illustrate the performance results, where each box represents the median and the interquartile range of the results. It is evident that for $\lambda = 0.1$, a better spread of results is obtained, indicating that sparse convolutive NMF achieves superior overall performance.

and used in 360 ($9 \times 4 \times 10$) separation experiments. For sparse convolutive NMF separation performance is tested for $\lambda = [0.01\ 0.1\ 0.3\ 1.0\ 2.0]$; resulting in 120 ($6 \times 4 \times 5$) speaker phone sets and 1800 ($9 \times 4 \times 5 \times 10$) separation experiments.

For the purposes of ease of comparison with existing separation methods, we evaluate the separation performance of both algorithms using the *source-to-distortion* ratio (SDR) measure provided by the `BSS_EVAL` toolbox [7]; SDR indicates overall separation performance and is expressed in dB, with higher performance values indicating better quality estimates. .

5.2 Separation Performance

We statistically analyse the performance of convolutive NMF and sparse convolutive NMF by collating the results from all experiments and presenting the results using box plots: Each box presents information about the median and the statistical dispersion of the results. The top and bottom of each box represents the upper and lower quartiles, while the length between them is the interquartile range; the whiskers represent the extent of the rest of the data, and outliers are represented by +. Box plots for SDR are presented in Figure 3.

The SDR results indicate that for $\lambda = [0.1, 0.3]$, the median performance obtained (0.66 dB, 0.62 dB) exceeds convolutive NMF (0.44 dB), for our given algorithm parameters. It is also evident that a better spread of results is produced for sparse convolutive NMF; demonstrating that when λ is chosen appropriately, sparse convolutive NMF achieves superior overall performance. However, audible reconstructions reveal that convolutive NMF is more resilient to artifacts; this may reflect the fact that each sparse phone set exhibits phones that are rich in features, which may manifest as artifacts in the resultant source estimates. It is

also evident that the performance of the sparse convolutive algorithm degrades significantly for large λ values, so much so, that it renders the results useless, for our data this is especially evident for $\lambda > 1$.

6 Conclusion

In this paper, we presented a sparse convolutive NMF algorithm, which effectively discovers a sparse parts-based representation for non-negative data. This method extends the convolutive NMF objective by including a sparseness constraint on the activation patterns, enabling the discovery of over-complete representations. Furthermore, we demonstrate the superiority of sparse convolutive NMF over convolutive NMF, when applied to a supervised monophonic speech separation task.

6.1 Acknowledgements

Supported by Higher Education Authority of Ireland (An tÚdarás Um Ard-Oideachas), and Science Foundation Ireland grant 00/PI.1/C067.

References

[1] P. Paatero and U. Tapper. Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–26, 1994.

[2] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Adv. in Neu. Info. Proc. Sys. 13*, pages 556–62. MIT Press, 2001.

[3] Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Fifth International Conference on Independent Component Analysis*, LNCS 3195, pages 494–9, Granada, Spain, September 22–24 2004. Springer-Verlag.

[4] S. A. Abdallah and M. D. Plumbley. Polyphonic transcription by non-negative sparse coding of power spectra. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, pages 318–25, 2004.

[5] Julian Eggert and Edgar Körner. Sparse coding and NMF. In *IEEE International Joint Conference on Neural Networks, 2004. Proceedings*, volume 4, pages 2529–2533. IEEE, July 2004.

[6] Paris Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE Transaction on Audio, Speech and Language Processing*, 2007.

[7] C Févotte, R Gribonval, and E Vincent. BSS_EVAL toolbox user guide. Technical Report 1706, IRISA, 2005.