NON-NEGATIVE MATRIX FACTORIZATION AS NOISE-ROBUST FEATURE EXTRACTOR FOR SPEECH RECOGNITION

Björn Schuller, Felix Weninger, Martin Wöllmer, Yang Sun, Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München
Arcisstrasse 21, D-80333 München, Germany
schuller@tum.de

ABSTRACT

We introduce a novel approach for noise-robust feature extraction in speech recognition, based on non-negative matrix factorization (NMF). While NMF has previously been used for speech denoising and speaker separation, we directly extract time-varying features from the NMF output. To this end we extend basic unsupervised NMF to a hybrid supervised/unsupervised algorithm. We present a Dynamic Bayesian Network (DBN) architecture that can exploit these features in a Tandem manner together with the maximum likelihood phoneme estimate of a bidirectional long short-term memory (BLSTM) recurrent neural network. We show that addition of NMF features to spelling recognition systems can increase word accuracy by up to 7 % absolute in a noisy car environment.

Index Terms— Non-Negative Matrix Factorization, Speech recognition, Noise robustness, Dynamic Bayesian Networks, Long Short-Term Memory

1. INTRODUCTION

Non-negative matrix factorization (NMF) and its extensions have been successfully used in areas related to speech recognition, including speech denoising and speaker separation [1–7]. The basic principle of NMF-based audio processing is to find a locally optimal factorization of a short-time magnitude spectrogram into two factors, of which the first one represents the spectra of the events occurring in the signal and the second one their time-varying gains. The mathematical background of NMF is explained in Sec. 2.

Previous works in NMF-based speech processing either aim for best separation quality or use NMF as a preprocessing step for conventional speech recognition procedures. In contrast, we propose to use the NMF algorithm as a data-based feature extractor. While a data-based NMF feature extraction process for sound classification has been described in [8], we aim at using NMF features as an addition to noise-robust speech recognition architectures. As an application scenario we chose in-car spelling recognition, where automatic speech recognition is especially useful as a hands-free intuitive human-machine interface. Here the speech recognition engine typically has to deal with negative signal-to-noise ratio (SNR) levels as well as similarly sounding utterances such as the letters *"b"* and *"d"* which are often hard to discriminate even for humans [9].

Our previous work in this area [10] has shown that a Tandem approach incorporating Dynamic Bayesian Networks (DBN) and the maximum likelihood phoneme index predicted by a bidirectional long short-term memory recurrent neural network (BLSTM) [11] performs well on this task. DBN-based architectures cannot only process features from the continous domain like the commonly used Mel frequency cepstral coefficients (MFCCs), but also discrete, value-restricted features such as the named index of the most likely phoneme per time frame. In this paper, we introduce another type of discrete feature based on NMF, which can accordingly be integrated into a DBN. To this end, we performed a supervised NMF variant with spectra that were pre-computed from spoken letter utterances to obtain the index of the component that contributes the most to the spectrum of each time frame.

The paper is structured as follows: first, we introduce the mathematical background of NMF and its usage for blind source separation in Sec. 2. Second, we describe our feature extraction procedure based on NMF in Sec. 3. Third, we describe the architecture of the Tandem DBN speech recognizer used to evaluate NMF feature extraction in Sec. 4. Finally, we show the results of our experiments with spelling sequences overlaid by in-car noise in Sec. 5 before concluding in Sec. 6.

2. NON-NEGATIVE MATRIX FACTORIZATION

2.1. Definition

Given a matrix $\mathbf{V} \in \mathbb{R}_+^{n \times m}$ and a constant $r \in \mathbb{N}$, non-negative matrix factorization (NMF) computes two matrices $\mathbf{W} \in \mathbb{R}_+^{n \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times m}$, such that

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

Usually one chooses $r \ll n, m$, so that NMF performs information reduction.

2.2. NMF in Signal Processing

NMF in signal processing is usually applied to magnitude spectra. Basic NMF approaches assume a linear signal model, i. e. that the short-time magnitude spectra of a monophonic signal can be expressed as linear combinations of spectra of several distinct components. Thereby the coefficients are restricted to be non-negative.

Considering Eq. 1, one can interpret the columns of $\mathbf{W}$ as spectral components and the corresponding rows of $\mathbf{H}$ as their time-varying gains. Assiging the spectral components to sources, this principle can be directly exploited for blind source separation.

In contrast, our work focuses on the extraction of features from the temporal structure revealed in the $\mathbf{H}$ matrix, and their usage for noise-robust automatic speech recognition.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

2.3. Factorization Algorithm

Factorization is usually achieved by iterative minimization of cost-functions. For our work, we choose the following function:

$$c_d(\mathbf{W},\mathbf{H}) = \sum_{ij}\left(\mathbf{V}_{ij}\log\frac{\mathbf{V}_{ij}}{(\mathbf{WH})_{ij}} - (\mathbf{V}-\mathbf{WH})_{ij}\right) \quad (2)$$

The function $c_d(\mathbf{W},\mathbf{H})$ has turned out to yield perceptually good results at a reasonable computational cost [12, 13], and it is the basis for several recent NMF-based techniques in speech processing [4–7].

While it can be extended to increase the perceived audio quality of components, this was of minor relevance for our work, which focuses on investigating whether NMF can be exploited for noise-robust feature extraction.

For minimization of (2), we implement Lee and Seung's algorithm [14] which iteratively modifies the matrices $\mathbf{W}$ and $\mathbf{H}$ using a 'multiplicative update'. It can be shown that throughout execution of this algorithm, the cost function (2) is *non-increasing* [14].

While $\mathbf{H}$ is initialized randomly, for $\mathbf{W}$ we use a 'targeted initialization' approach which will be explained in the next section.

3. NMF FEATURE EXTRACTION

3.1. Supervised Variant of NMF

It is characteristic for speech-related tasks that prior knowledge about the events in the signal is available, thus a matrix $\mathbf{W}$ can be predefined. For example, for our spelling recognition scenario, we compute a matrix $\mathbf{W}$ whose columns contain spectra of spelled letters. Throughout the iteration $\mathbf{W}$ is kept constant whereas $\mathbf{H}$ is updated iteratively.

This variant of NMF is a *supervised* algorithm that finds a representation of the signal using the columns in $\mathbf{W}$ as basis vectors. The computed matrix $\mathbf{H}$ can be interpreted as time-dependent values of r different features of the signal that can be used as input for a dynamic classifier, e. g. a DBN or recurrent neural net.

The calculation of a $\mathbf{W}$ matrix for supervised NMF can be summarized as follows: for each event (e. g. letter) $e \in \{1,\ldots,E\}$:

1. Concatenate the corresponding training samples
2. Compute the magnitude spectrogram $\mathbf{V}_e$ by short-time Fourier transformation
3. From $\mathbf{V}_e$ obtain matrices $\mathbf{W}_e$, $\mathbf{H}_e$ by NMF

Note that each $\mathbf{W_e}$ contains 'characteristic' spectra of event e, i. e. the spectra that model all of the training samples with the least overall error. From the $\mathbf{W_e}$ we build the matrix $\mathbf{W}$ by column-wise concatenation:

$$\mathbf{W} := \mathbf{W}_1|\mathbf{W}_2|\cdots|\mathbf{W}_E,$$

A similar technique has been used in [4, 5], aiming at separation of speech and noise, and in [1, 2] for speaker separation.

3.2. Hybrid Supervised / Unsupervised Approach

For scenarios where events in the signal are overlaid with unknown noise, we combine the supervised NMF variant introduced in the previous section with the conventional NMF algorithm [14]. To this end, we enhance the $\mathbf{W}$ matrix containing pre-computed spectra with one or more randomly initialized 'noise' columns. Only these columns are updated in each iteration step, using the $\mathbf{W}$ update rule from [14].

Intuitively speaking, this algorithm finds a signal representation using fixed basis vectors, putting everything that cannot be described with these vectors into a noise component. In our work, the gains of this component were considered irrelevant for the recognition task.

3.3. Reduction of the Feature Space

The aforementioned methods generate a large number of highly correlated features. It is therefore advisable to reduce the feature space. While this could be achieved by methods like PCA or LDA that aim at preservation of all original feature information, we found that the information contained in the component with the highest gain was sufficient for our purpose.

Thus, for each time frame t we calculate the discrete feature g_t:

$$g_t = \arg\max_i(\mathbf{H}_{it}), i \in \{1,\ldots,R\} \quad (3)$$

where R is the number of NMF components with pre-initialized spectra. This feature will subsequently be referred to as 'NMF maximum gain component index'.

4. DBN DECODERS WITH BLSTM AND NMF FEATURES

The DBN architecture processing BLSTM and NMF features is depicted in Fig. 1. The lower, grey-shaded part of the figure shows the basic neural network architecture of the BLSTM net, consisting of an input $\mathbf{i}_t$, an output $\mathbf{o}_t$, and a hidden node h_t for each time step. Details on the principle of Long Short-Term Memory [15] networks can be found in [11] or [10]. The upper part of Fig. 1 shows the explicit DBN representation of a Hidden Markov Model following the DBN structure as introduced in [16].

Similarly to [16] or [10], for every time step, the following random variables are defined: w_t represents the current word, w_t^{ps} denotes the position within the word, w_t^{tr} is a binary indicator variable for a word transition, and s_t is the HMM state with s_t^{tr} indicating a state transition. The variable $\mathbf{x}_t$ denotes the observed acoustic features. b_t and g_t respectively contain the phoneme prediction of the BLSTM (as in [10]) and the index of the NMF component with the highest gain (Eq. 3) as additional discrete observations. The size of the BLSTM input layer $\mathbf{i}_t$ corresponds to the dimensionality of the acoustic feature vector $\mathbf{x}_t$ whereas the vector $\mathbf{o}_t$ contains one probability score for each of the P different phonemes at each time step. b_t is the index of the maximum likelihood phoneme:

$$b_t = \arg\max_j(\mathbf{o}_{t,j}), j \in \{1,\ldots,P\} \quad (4)$$

The DBN structure in Fig. 1 displays hidden variables as circles and observed variables as squares. Straight lines represent deterministic conditional probability functions (CPFs) whereas random CPFs correspond to zig-zagged lines. Dotted lines refer to so-called *switching parents* which in our case switch between two different CPFs as in [16]. The CPFs $p(b_t|s_t)$, $p(g_t|s_t)$ and $p(s_t^{tr}|s_t)$ are probability tables that are learnt via EM training.

An alternative architecture is shown in Fig. 2. Here the acoustic feature vector is modelled by a CPF $p(x_t|s_t)$ which is described by a Gaussian mixture as common in an HMM system. One could also consider a combination of these approaches, i. e. let the model in Fig. 1 also process the MFCC acoustic features. However, in our experiments this did not produce significantly better results.

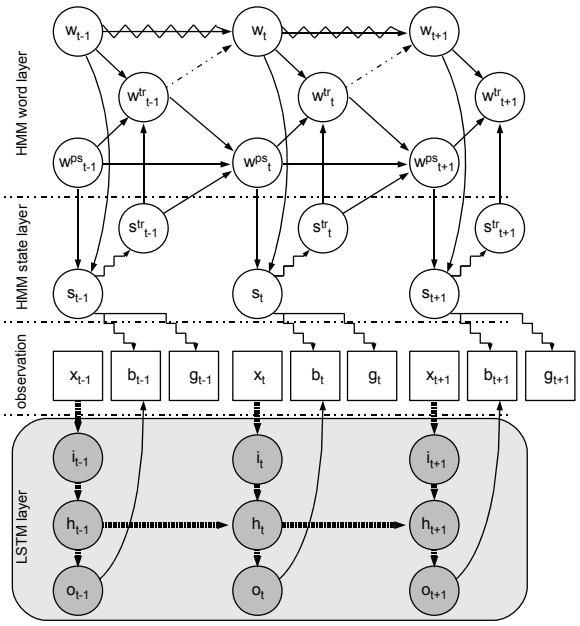

Fig. 1. Architecture of a DBN processing the BLSTM maximum likelihood phoneme index b_t and NMF maximum gain component index g_t. Note that the MFCC acoustic features $\mathbf{x}_t$ are not directly used by the DBN, but rather given as input to the BLSTM.

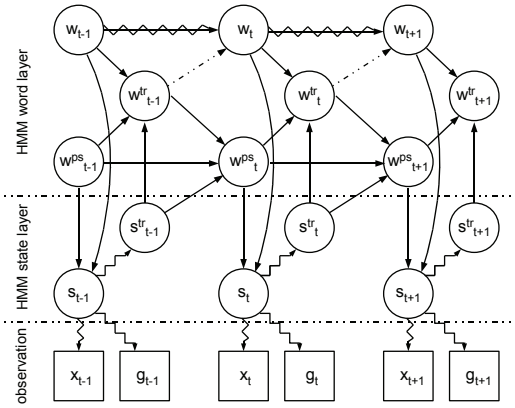

Fig. 2. Architecture of a DBN with MFCC features $\mathbf{x}_t$ and NMF maximum gain component index g_t. Note that here the vector $\mathbf{x}_t$ is directly processed by the DBN.

5. EXPERIMENTS

For the evaluation of the NMF feature extraction, we chose the task of noise-robust in-car spelling recognition. We used the letter utterances from *"a"* to *"z"* from the TI 46 Speaker Dependent Isolated Word Corpus to generate a large set of spelling sequences. A detailed description of the database can be found in [10]. Out of the clean spelling utterances, noisy sequences were generated by superposing the speech signal with different in-car noise types as used e. g. in [9].

Since the road surface has a strong influence on the characteristics of in-car noise, three different surfaces in combination with typical velocities have been considered: driving on a smooth city road (CTY), driving on a highway (HWY), and driving on cobbles (COB). The resulting SNR histogram of the noisy speech utterances correspond to the one shown in [10].

We considered the speech recognition architectures depicted in Fig. 1 and 2, both with and without the NMF maximum gain component index g_t. MFCC feature extraction, as well as the letter HMMs and the BLSTM network correspond exactly to the experiments in [10].

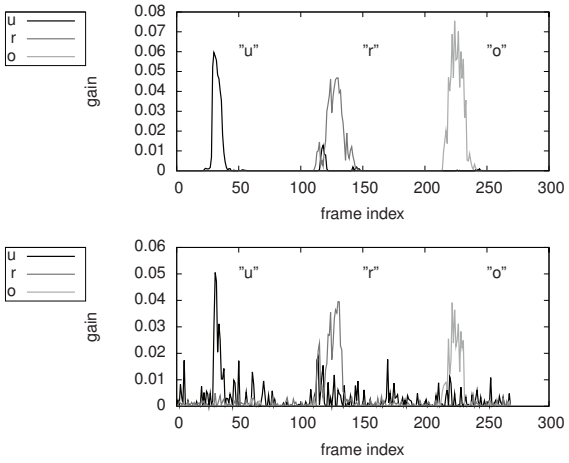

Fig. 3. NMF gains over time for the spoken letter sequence *"u r o"*. 3 exemplary components are shown for the clean (top) and COB noise (bottom) cases. Notably, the shape of the gains is similar in the clean and noisy cases.

NMF maximum gain component indices were computed as follows: for every letter from *"a"* to *"z"* we created 26 signals by concatenating the utterances from all speakers in the original TI46 training set. We computed the short-time Fourier spectrograms of these signals, using a Hamming window, 10 ms frame rate and 25 ms window size. Then we applied three-component NMF to the spectrograms, yielding $26 \times 3 = 78$ spectra.

We used the 78 component spectra to perform supervised NMF with 78 components on each of the clean signals in the training and test set of our database. NMF on the noisy sequences (training and test sets) was performed using the hybrid supervised/unsupervised approach, using one additional noise component, thus 79 components in total. Best results were achieved when the noisy sequences were not pre-filtered before NMF. For each time frame the index of the component with the highest gain was computed, neglecting the 79^{th} (noise) component for the noisy sequences.

Fig. 3 shows the 'raw' gains (i. e. entries of $\mathbf{H}$) over time for three exemplary components, computed by applying the aforementioned supervised NMF procedure to the letter sequence *"u r o"*, spoken by male speaker 8 from the TI46 database. This sequence is pronounced [Y UW . AA R . OW] in CMU notation. As can be determined empirically, the shown components roughly correspond to the phonemes [Y] (component from the letter *"u"*), [AA] (from letter *"r"*), and [AO] " (letter *"o"*).

Comparing the upper plot (clean case) and the lower plot (noisy case), one notices a small difference in scaling. Because in the noisy case an additional component was introduced, the NMF can not only model noise with it, but also differences in the letter spectra between training and test set. Thus the pre-defined component spectra receive a smaller share of the total gain. Note that scaling has no influence on the maximum gains component index. Furthermore, due to the noise,

some components have non-zero gains during the silence between letters. Apart from that, the shapes of the graphs look rather similar.

model	test cond.	M	M+N	B	B+N
clean	clean	98.19	98.52	92.14	96.39
CTY	CTY	92.64	94.57	90.33	96.14
HWY	HWY	84.06	88.77	87.61	91.57
COB	COB	81.65	86.79	87.91	92.69
CTY	HWY	60.50	80.62	82.28	95.50
CTY	COB	64.38	72.96	80.64	95.48
HWY	CTY	54.25	56.38	87.80	91.51
HWY	COB	59.09	58.63	84.65	92.03
COB	CTY	79.07	71.72	86.43	92.70
COB	HWY	74.32	80.06	85.08	92.42
	mean	**74.82**	**78.90**	**86.70**	**93.64**

Table 1. Spelling recognition word accuracies in percent for DBN with MFCC features (M), MFCCs and NMF feature (M+N), BLSTM feature (B) and BLSTM+NMF features (B+N), matched and mismatched condition

Tab. 1 shows the word accuracies (WA) for a DBN with 39 MFCC features as described above (M), a DBN with MFCC features plus NMF maximum gain component index g_t (M+N, as depicted in Fig. 2), a DBN with the BLSTM maximum likelihood phoneme index b_t (B) and a DBN with both b_t and g_t (B+N, see Fig. 1). Going from left to right the improvement of the mean WA is statistically significant at a level of 10^{-3} throughout using a one-tailed t-test. The upper half contains the 'matched condition', the lower half the 'mismatched condition' cases. Note that a model trained on perfectly clean data fails in noisy test conditions since the silence model will tolerate no signal variance at all, which would lead to permanent insertion errors. In clean conditions performance is only slightly enhanced by the NMF component index. As soon as the speech signal is corrupted by noise, performance decreases whereas in the matched condition case the NMF component index increases performance by up to 5 % absolute. For the mismatched condition case, the greatest improvement can be observed for a model trained on a smooth inner city road (CTY) and tested on the highway (HWY). There, the NMF component index can increase word accuracy by over 20 % absolute. However, for the model trained on a cobble road (COB) and tested on a smooth inner city road (CTY), a decrease in performance by about 7 % absolute is observed.

Yet, although the DBN with BLSTM maximum likelihood phoneme index already performs very well in terms of word accuracy, addition of the NMF maximum gain component index yields a further improvement of up to 15 % absolute (in the CTY model / COB test case), and by 7 % absolute on average over all test conditions.

6. CONCLUSION

We introduced a novel type of feature for speech recognition based on the results of supervised NMF. We showed that on the one hand, this feature can significantly improve the recognition rate of a DBN that uses traditional acoustic features. On the other hand, a DBN combining the NMF feature with the phoneme prediction of a BLSTM recurrent neural net produced the best results in terms of word accuracy, which is raised over 93 %.

Future work will investigate whether the variety of enhanced NMF algorithms that improve source separation quality can also be exploited for better NMF feature extraction. Also, we want to investigate usefulness of the proposed NMF feature extraction metaphor in related audio, speech, and music processing tasks, especially large-vocabulary ASR.

7. REFERENCES

[1] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non- negative matrix factorization," in *Proc. of Interspeech*, Pittsburgh, Pennsylvania, 2006.

[2] P. D. O'Grady and B. A. Pearlmutter, "Discovering convolutive speech phones using sparseness and non-negativity constraints," in *Proc. of ICA*, London, UK, 2007.

[3] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Efficient model-based speech separation and denoising using non-negative subspace analysis," in *Proc. of ICASSP*, Las Vegas, USA, 2008.

[4] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. of ICASSP*, Las Vegas, USA, 2008.

[5] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Proc. of Interspeech*, Brisbane, Australia, 2008.

[6] T. Virtanen and A. T. Cemgil, "Mixtures of gamma priors for non-negative matrix factorization based speech separation," in *Proc. of ICA*, Paraty, Brazil, 2009.

[7] T. Virtanen, "Spectral covariance in prior distributions of non-negative matrix factorization based speech separation," in *Proc. of EUSIPCO*, Glasgow, Scotland, 2009.

[8] Y.-C. Cho, S. Choi, and S.-Y. Bang, "Non-negative component parts of sound for classification," in *Proc. of ISSPIT*, Darmstadt, Germany, 2003, pp. 633–636.

[9] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Recognition of noisy speech: A comparative survey of robust model architectures and feature enhancement," *EURASIP Journal on Audio, Speech, and Music Processing*, 2009, id 942617.

[10] M. Wöllmer, F. Eyben, B. Schuller, Y. Sun, T. Moosmayr, and N. Nguyen-Thien, "Robust in-car spelling recognition - A Tandem BLSTM-HMM approach," in *Proc. of Interspeech*, Brighton, UK, 2009.

[11] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. of ICANN*, Warsaw, Poland, 2005, pp. 602–610.

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

[13] B. Schuller, A. Lehmann, F. Weninger, F. Eyben, and G. Rigoll, "Blind enhancement of the rhythmic and harmonic sections by NMF: Does it help?," in *Proc. of the International Conference on Acoustics (NAG/DAGA 2009)*, Rotterdam, The Netherlands, 2009.

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. of NIPS*, Vancouver, Canada, 2001, pp. 556–562.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9(8), pp. 1735–1780, 1997.

[16] J. A. Bilmes and C. Bartels, "Graphical model architectures for speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 89–100, 2005.