Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization

Björn Schuller, Felix Weninger

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, and Felix Weninger. 2010. "Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization." In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 14-19 March 2010, Dallas, TX, USA*, edited by Scott C. Douglas and Nasser Kehtarnavaz, 5054–57. Piscataway, NJ: IEEE. https://doi.org/10.1109/ICASSP.2010.5495061.

DISCRIMINATION OF SPEECH AND NON-LINGUISTIC VOCALIZATIONS BY NON-NEGATIVE MATRIX FACTORIZATION

Björn Schuller and Felix Weninger

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

ABSTRACT

We introduce features based on Non-Negative Matrix Factorization (NMF) for discrimination of speech and non-linguistic vocalizations such as laughter or breathing, which is a crucial task in recognition of spontaneous speech. NMF has been successfully used in speech-related tasks such as de-noising and speaker separation. While existing approaches use it as a preprocessing step for conventional speech recognizers, we aim at directly classifying the output of the NMF algorithm. To this end, we propose a feature extraction procedure based on a supervised variant of NMF, considering two different algorithms. Applying our approach to a spontaneous speech corpus, we show that addition of NMF features to an MFCC-based classifier increases mean recall of speech and non-linguistic vocalizations by over 2.5 % absolute, and particularly recall of laughter by 6.6 % absolute. The improvement is significant at a level of 0.4 %.

Index Terms— Non-Negative Matrix Factorization, Non-linguistic vocalizations, Speech recognition, Spontaneous speech

1. INTRODUCTION

Non-negative matrix factorization (NMF) and its extensions have been successfully used in areas related to speech recognition, including speech de-noising and speaker separation [1–6].

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 activation over time. 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 signal enhancement technique that is applied before 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 [7], we aim at using NMF features as an addition to conventional acoustic features for discrimination of speech and non-linguistic vocalizations, including laughter, breathing, hesitation (e. g. *"uhm"*) or non-verbal consent (e. g. *"aha"*). To this end, we perform a supervised NMF variant with spectra that were pre-computed from instances of speech and non-linguistic vocalizations to measure which spectra contribute the most to the signal.

Discrimination of speech and non-linguistic vocalizations plays an important role in speech recognition systems dealing with spontaneous speech, like dialog systems, call center loops or automatic transcription of meetings. In contrast to read speech, which conveys only the information contained in the spoken words and sentences, spontaneous speech contains more extra linguistic information. As non-linguistic vocalizations reveal much about this information [8], it is vital for a spontaneous speech recognizer to spot non-linguistic vocalizations and their type [9].

Several specialized approaches have been proposed for the detection of filled pauses [10] and laughter [11–13]. In contrast, our previous work [14] in this area considered the discrimination of five types of non-linguistic vocalizations in a purely data-driven manner. Experiments were carried out on segments extracted from the Audio-Visual Interest Corpus (AVIC) [15]. In this paper, we extend this approach to not only distinguish between different types of non-linguistic vocalizations, but also to discriminate them from speech. Furthermore, we will aim at a speaker-independent recognizer, and most notably we will show that features generated by NMF can increase classification accuracy compared to traditional acoustic features such as Mel frequency cepstral coefficients (MFCCs). Thereby we also consider the impact on accuracy caused by different NMF algorithms.

The paper is structured as follows: first, we introduce the mathematical background of NMF and its usage in signal processing in Sec. 2. Second, we describe our feature extraction procedure based on NMF in Sec. 3. Third, we show the results of our experiments with speech and non-linguistic vocalizations from the AVIC corpus in Sec. 4 before concluding in Sec. 5.

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.

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).

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 activations. In particular, the i-th row of the $\mathbf{H}$ matrix indicates the amount that the spectrum in the i-th column of $\mathbf{W}$ contributes to the spectrogram of the original signal. This fact is the basis for our feature extraction approach, which will be explained in Sec. 3. But first, the algorithmic aspects of NMF shall be discussed.

2.3. Factorization Algorithm

Factorization is usually achieved by iterative minimization of cost-functions. An 'obvious' cost-function is the squared Euclidean distance between the original matrix and the product of the NMF factors:

$$c_e(\mathbf{W},\mathbf{H}) = ||\mathbf{V}-\mathbf{WH}||_F = \sum_{i=1}^{n}\sum_{j=1}^{m}(\mathbf{V}-\mathbf{WH})_{ij}^2 \quad (2)$$

This function is the basis of an approach for speaker separation with NMF [1]. However, several recent works in NMF-based speech processing [3–6] use cost-functions based on a modified version of Kullback-Leibler (KL) divergence:

$$c_d(\mathbf{W},\mathbf{H}) = \sum_{ij}\left(\mathbf{V}_{ij}\log\frac{\mathbf{V}_{ij}}{(\mathbf{WH})_{ij}} - (\mathbf{V}-\mathbf{WH})_{ij}\right) \quad (3)$$

Thus, both c_d and c_e have been successfully used for NMF in speech-related tasks, but to our knowledge the impact of different cost-functions on the results has not been thoroughly evaluated apart from the field of blind source separation in music [16]. As the feature extraction procedure in Sec. 3 is independent from the NMF cost-function, we could use either c_d or c_e in our experiments and show a comparison of both in Sec. 4.

For minimization of either cost-function, we implemented the two algorithms by Lee and Seung [17], which iteratively modify $\mathbf{W}$ and $\mathbf{H}$ using 'multiplicative update' rules.

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

We now consider a supervised variant of NMF using a predefined matrix $\mathbf{W}$ which is kept constant throughout the iteration while $\mathbf{H}$ is updated iteratively. In this case, NMF seeks a minimal-error representation of the signal (in terms of the cost-function) with only a set of given spectra.

As outlined in Sec. 2.2, the $\mathbf{H}$ matrix measures the contribution of spectra to the original signal. Thus, by using a matrix $\mathbf{W}$ that contains spectra of speech and different non-linguistic vocalizations, the rows of $\mathbf{H}$ provide information whether the original signal consists of speech or a certain type of non-linguistic vocalization.

Assuming that we aim at discrimination of C different classes of signals, our algorithm for the computation of a $\mathbf{W}$ matrix for supervised NMF can be summarized as follows: for each class $c \in \{1,\ldots,C\}$:

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

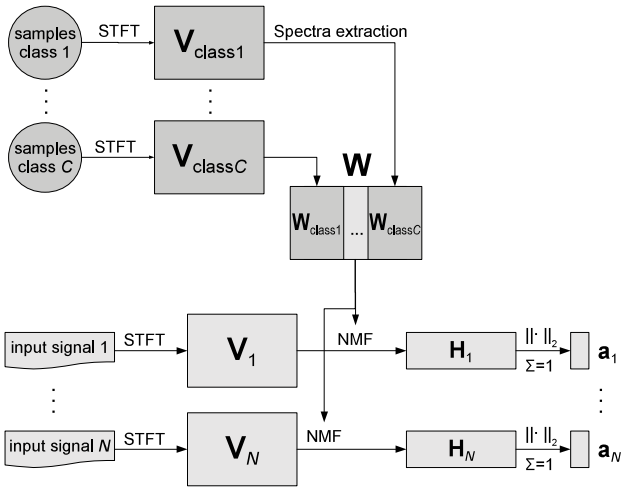

Fig. 1. Block diagram showing the extraction of NMF activation features for discrimination of C classes in N input signals. Matrices denoted by $\mathbf{V}$ are spectrograms. The matrix $\mathbf{W}$ consists of spectra computed from training data according to Sec. 3.1 and is used to perform supervised NMF on the spectrograms of all input signals. Feature extraction is carried out on the resulting $\mathbf{H}$ (activation) matrices. $||\cdot||_2$ indicates that the Euclidean norm of each matrix row is computed, and $\sum = 1$ is a normalization such that the components of each vector $\mathbf{a}_i$ sum up to 1.

Intuitively speaking, each $\mathbf{W}_c$ contains 'characteristic' spectra of class c. More precisely, these are the spectra that model all of the training samples belonging to class c with the least overall error. From the $\mathbf{W}_c$ we build the matrix $\mathbf{W}$ by column-wise concatenation:

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

In our experiments, we empirically determined the number of components to use for each class. Best results were achieved with 10-20 components per class (see Sec. 4 for details).

Note that a similar technique has been used e. g. in [3, 4], aiming at separation of speech and noise, and in [1] for speaker separation.

3.2. NMF Activation Features

From $\mathbf{H} \in \mathbb{R}^{r\times m}$ we calculate a feature vector $\mathbf{a} \in \mathbb{R}^r$ such that $\mathbf{a}_i$ is the Euclidean length of the i-th row of $\mathbf{H}$. To obtain features that are independent of the length and power of the signal, we normalize $\mathbf{a}$ such that $|\mathbf{a}|_1 = 1$.

The components of the vector $\mathbf{a}$ shall be subsequently referred to as 'NMF activation features'. A block diagram summarizing the procedure presented in this section is given in Fig. 1.

4. EXPERIMENTS

4.1. Data Set

We prepared a data set based on the Audio-Visual Interest Corpus (AVIC). This corpus consists of 3 901 turns, spoken by 21 subjects (10 of them female). The total recording time for males resembles 5:14:30 h with 1 907 turns, for females 5:08:00 h with 1 994 turns,

respectively. The spoken content, including non-linguistic vocalizations, is transcribed on the word level. For a detailed description of AVIC we refer to [15].

We divided the corpus into a training, development, and test set. Each of the 21 speakers was assigned to exactly one of the sets to evaluate the recognition procedure in a speaker-independent manner. The sets were chosen such that the total length of the utterances is approximately equal for each set, and furthermore each set is balanced by the length of male and female utterances, as is the whole corpus. In detail, speakers 4, 6, 8, 10, 28, 34 and 36 were assigned to the training set, speakers 12, 13, 15, 26, 30, 33, 35 to the development set, and speakers 5, 9, 16, 27, 29, 31, 32 to the test set.

We used the transcription of the corpus to extract the signal parts containing non-linguistic vocalizations of the 4 classes 'consent', 'laughter', 'hesitation', and 'breathing'. Note that we ignored the 'coughing' class due to a very small number of instances, as well as the 'garbage' class corresponding to other human noise, as it is rather application dependant. Then we generated 'speech-only' turns by cutting out all segments that contain non-linguistic vocalizations according to the transcription. Thus we ended up with instances from 5 classes. We eliminated all instances with a length of less than 100 ms, as feature extraction and model evaluation on such short segments is very error-prone [14]. In total, the training set consisted of 2 070, the development set of 1 980 and the test set of 2 184 instances.

As the number of instances of the 'speech' class is clearly dominating, we applied upsampling to the training set (when testing with the development set) and to the union of training and development set (when testing with the test set). Thereby all instances of each of the 'smaller' classes were duplicated such that these classes had roughly equal numbers of instances.

4.2. Feature Extraction and Classification

Each instance was transferred to the frequency domain by applying STFT with a Hamming window, 25 ms window size, and 10 ms frame rate. From the resulting spectrograms, we extracted Mel frequency cepstral coefficients (MFCCs) 0-12 with 26 filter banks, as common in speech processing. We considered the mean and standard deviation as time- and length-independent functionals of each MFCC. Furthermore we added the MFCCs from 5 equidistant signal frames, starting with the first and ending with the last signal frame. Finally, we also computed the first-order (δ) and second-order regression ($\delta\delta$) coefficients, and added their mean and standard deviation as well as their values at 5 equidistant signal frames, yielding a total of 273 acoustic features per instance.

Performance of the NMF activation features was evaluated by training with the training set and testing with the development set. Spectra for the extraction of NMF activation features were computed from signals concatenated from all the training utterances for each of the laughter, consent, hesitation, and breathing classes, respectively. Note that for this concatenation the original training set (before down sampling) was used, as from our experience the performance of NMF features increases when spectra are computed from longer input signals. However, for the speech class, as in [1], only 10 % of the training material was used, since a factorization of all speech utterances from the training set is not feasible considering memory requirements and limitations. The NMF cost-function as well as the number of components was varied, as will be explained in the next section.

Each feature is linearly scaled to the range $[-1, 1]$. As classifier, we used Support Vector Machines (SVM). It turned out that in our task SVM with radial basis functions (RBF) outperform SVM with linear kernel.

Recall [%]	**Euclidean distance**			**mod. KL divergence**		
	N70	**N90**	**N100**	**N70**	**N90**	**N100**
speech	66.69	68.35	69.65	69.50	**71.23**	71.02
hesitation	67.63	61.35	62.80	71.01	73.19	**76.09**
consent	84.18	86.44	**89.27**	**89.27**	88.70	85.88
laughter	51.32	55.26	47.37	71.05	**75.00**	71.05
breathing	**95.38**	91.54	94.62	87.69	88.46	91.54
mean	73.04	72.59	72.73	77.71	**79.32**	79.11

Table 1. Recall and mean recall for 4 different non-linguistic vocalizations and speech on the test set of the AVIC corpus, consisting of 2 184 utterances from 4 male and 3 female subjects. All results are achieved using an SVM with RBF kernel, trained with different sets of NMF activation features: N70, N90, and N100, corresponding to 70, 90, and 100 NMF components, respectively. NMF was computed by minimization of either Euclidean distance (Eq. 2) or modified KL divergence (Eq. 3). The best result per class is highlighted.

4.3. Results

After combining the training and development set, we evaluated the performance of the aforementioned feature extraction and classification procedure on the test set. To this end, we extracted NMF activation features using spectra that were computed from a concatenation of samples from both the training and development set.

We considered NMF activation features computed with 70, 90, and 100 components (N70, N90, N100). The spectra for 70 components were distributed among the classes as follows: 20 for the speech and laughter classes, and 10 for the remaining three classes (consent, hesitation, breathing), corresponding to the fact that speech and laughter are probably more diverse than hesitation, consent, or breathing. For 90 components, the number of speech spectra was doubled from 20 to 40; for 100 components, 20 spectra for each of the 5 classes were used. Both Euclidean distance (Eq. 2) and modified KL divergence (Eq. 3) were considered as cost-functions.

The results achieved with these features are shown in Table 1, indicating the recall in percent for an SVM classifier with RBF kernel. As to the cost-function, we conclude that NMF feature extraction works best when minimizing modified KL divergence, outperforming Euclidean distance by up to 7 % absolute. The improvement observed by choosing modified KL divergence over Euclidean distance for computing the N100 feature set is significant at $p \approx 1.1 \cdot 10^{-5}$ (one-tailed McNemar test). Particularly the Euclidean distance features perform poorly in the detection of laughter, which is mostly misclassified as breathing. Compared to the choice of the cost-function, the number of NMF components seems to have a smaller influence on performance. Overall the best mean recall is achieved by using 90 components and minimizing modified KL divergence.

Results achieved by using MFCC features and MFCC together with NMF activation features are shown in Table 2. The columns denote the recall in percent for an SVM classifier with RBF kernel, provided with 273 MFCC features (M) or both MFCC and one of the NMF activation feature sets (M+N70, M+N90, M+N100). For the experiments in this table, NMF activation features were computed by minimization of Euclidean distance, which yielded slightly – yet not significantly – better results.

First, it is evident that, though performing considerably well in

Recall [%]	M	M+N70	M+N90	M+N100
speech	79.52	80.25	80.61	**80.97**
hesitation	85.99	**86.96**	85.99	85.27
consent	89.27	90.96	92.09	**93.79**
laughter	81.58	86.84	86.84	**88.16**
breathing	93.85	93.85	**94.62**	**94.62**
mean	86.04	87.77	88.03	**88.56**

Table 2. Recall and mean recall for 4 different non-linguistic vocalizations and speech on the test set of the AVIC corpus, consisting of 2 184 utterances from 4 male and 3 female subjects. All results are achieved using an SVM with RBF kernel. The columns indicate four different feature sets: 273 MFCC features (M) and MFCC plus NMF activation features (M+N70, M+N90, M+N100), corresponding to 70, 90, and 100 NMF components as in Table 1. NMF activation features are computed by minimization of Euclidean distance (Eq. 2). The best result per class is highlighted.

detection of breathing and consent, NMF activation features alone cannot surpass MFCCs in terms of mean recall. However, considering the impact of adding NMF activation features to MFCC features, it can be seen that they increase the recall for all classes but hesitation.

Notably the M+N100 set yields the best mean recall and drastically increases the recall for the laughter class by 6.6 % absolute, compared to MFCC features only. For the consent class there is an increase of 4.5 % absolute. Conducting the one-tailed McNemar test reveals that the improvement by using the M+N100 feature set instead of MFCCs only (M) is significant at $p \approx 3.76 \cdot 10^{-3}$.

5. CONCLUSION

We presented a supervised NMF procedure to compute 'activation features' and could show that they perform considerably well for the discrimination of speech and non-linguistic vocalizations. Thereby an NMF algorithm minimizing a criterion related to Kullback-Leibler divergence produced clearly better results than the corresponding algorithm for Euclidean distance. Furthermore, we demonstrated that this type of feature can significantly improve performance of MFCC-based static classifiers: in an experiment carried out on a large corpus of spontaneous speech, NMF activation features could increase unweighted mean recall by 2.5 % absolute, recall of the 'consent' type of vocalization by 4.5 % absolute, and recall of laughter by 6.6 % absolute.

Our future work in this area will investigate whether extensions of NMF, such as non-negative matrix deconvolution [18], or various extensions of the cost-functions such as sparseness constraints [1] can further improve the performance of NMF features. Furthermore, we want to include a greater variety of audio features such as the ones proposed in [14] and extend the classification procedure by decorrelation and feature selection techniques. Finally, our goal is to extend the presented feature extraction metaphor to use time-varying features with dynamic classifiers for segmentation of signals into speech and non-linguistic vocalizations.

6. 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] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals," in *Proc. of Interspeech*, Antwerp, Belgium, 2007, pp. 2253–2256.

[9] N. Campbell, "On the use of nonverbal speech sounds in human communication," in *Proc. of COST 2102 Workshop*, Vietri sul Mare, Italy, 2007, pp. 117–128.

[10] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system for spontaneous speech recognition," in *Proc. of Eurospeech*, Budapest, Hungary, 1999, pp. 227–230.

[11] K. P. Truong and D. A. van Leeuwen, "Automatic detection of laughter," in *Proc. of Interspeech*, Lisbon, Portugal, 2005, pp. 485–488.

[12] N. Campbell, H. Kashioka, and R. Ohara, "No laughing matter," in *Proc. of Interspeech*, Lisbon, Portugal, 2005, pp. 465–468.

[13] M. Knox and M. Mirghafori, "Automatic laughter detection using neural networks," in *Proc. of Interspeech*, Antwerp, Belgium, 2007.

[14] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," in *Perception in Multimodal Dialogue Systems, Proc. of PIT 2008*, pp. 99–110. Springer, 2008.

[15] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing Journal*, 2009.

[16] 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.

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

[18] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," in *Proc. of SAPA*, Jeju, Korea, 2004.