UNSUPERVISED LEARNING OF AUDITORY FILTER BANKS USING NON-NEGATIVE MATRIX FACTORISATION

Alexander Bertrand, Kris Demuynck, Veronique Stouten, Hugo Van hamme

Katholieke Universiteit Leuven - Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

ABSTRACT

Non-negative matrix factorisation (NMF) is an unsupervised learning technique that decomposes a non-negative data matrix into a product of two lower rank non-negative matrices. The non-negativity constraint results in a parts-based and often sparse representation of the data. We use NMF to factorise a matrix with spectral slices of continuous speech to automatically find a feature set for speech recognition. The resulting decomposition yields a filter bank design with remarkable similarities to perceptually motivated designs, supporting the hypothesis that human hearing and speech production are well matched to each other. We point out that the divergence cost criterion used by NMF is linearly dependent on energy, which may influence the design. We will however argue that this does not significantly affect the interpretation of our results. Furthermore, we compare our filter bank with several hearing models found in literature. Evaluating the filter bank for speech recognition shows that the same recognition performance is achieved as with classical MEL-based features.

Index Terms— Non-negative matrix decomposition, Unsupervised learning, Speech analysis, Feature extraction, Auditory system

1. INTRODUCTION

The goal of this work is to automatically discover a feature set for speech recognition by analysing continuous speech recordings, using only minimal knowledge about audiology or phonology. We perform a non-negative matrix factorisation (NMF) of a data matrix containing power spectra of continuous speech. Unlike classical factorisation techniques like PCA, NMF generates a parts-based representation of the data. There is psychological and physiological evidence that human perception is often based on such representation [1]. The individual parts found by NMF are usually sparse and must be combined in an additive (not subtractive) linear combination.

Smaragdis [2] extended NMF to a technique known as 'convolutive NMF' to find 2D spectro-temporal objects in speech spectrograms. These objects seem to be roughly related to phone instances of speech. Lewicki [3] used time-domain ICA without non-negativity constraints on short segments of natural sounds (including speech) to find the underlying structure. The resulting filters show a trade-off between localisation in time or in frequency, which is qualitatively similar to the response properties of auditory fibers.

Our goal is to discover a small featureset which can immediately be plugged into state-of-the-art speech recognition systems. In contrast to the works stated above, we therefore ignore the temporal characteristics of speech and aim for a much higher dimensionality reduction. We will compare our result with the perceptually motivated MEL-based features which are widely used in speech recognition systems.

2. NON-NEGATIVE MATRIX FACTORISATION USING THE DIVERGENCE CRITERION

NMF decomposes a non-negative data matrix $\mathbf{V}$ into a product of two lower rank non-negative matrices $\mathbf{W}$ and $\mathbf{H}$:

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

with $\mathbf{V}$ an $m \times n$ matrix, $\mathbf{W}$ an $m \times r$ matrix and $\mathbf{H}$ an $r \times n$ matrix where normally $r \leq m$. This shows that each column of $\mathbf{V}$ is written as a linear combination of the r basis vectors in the columns of $\mathbf{W}$, weighted with the coefficients in the corresponding column of $\mathbf{H}$. This can be seen as a dimensionality reduction of data vectors in an m-dimensional space to the r-dimensional space spanned by the columns of $\mathbf{W}$. This is only possible if the basis vectors represented by the columns of $\mathbf{W}$ uncover the latent linear structure in the data.

The quality of the approximation can be assessed with cost functions such as mean squared error (MSE) or divergence [4]. We focus only on the divergence criterion because the dynamic range of the data used in this experiment is too high to use the MSE criterion (see section 3). The divergence criterion between matrices $\mathbf{V}$ and $\mathbf{X}$ is defined as

$$Div(\mathbf{V} \parallel \mathbf{X}) = \sum_{i,j} \left(\mathbf{V}_{ij} \log \frac{\mathbf{V}_{ij}}{\mathbf{X}_{ij}} - \mathbf{V}_{ij} + \mathbf{X}_{ij} \right) \tag{2}$$

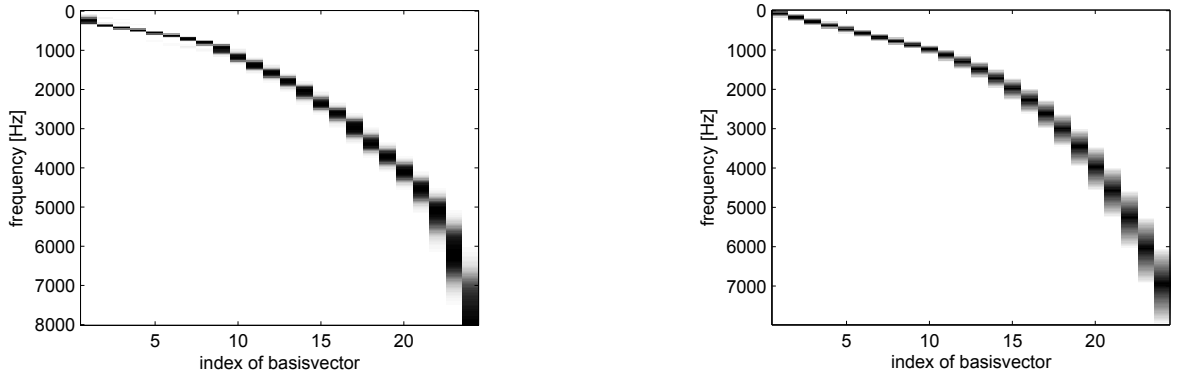
which reduces to zero if and only if $\mathbf{V} = \mathbf{X}$.

To find stationary points of the divergence between $\mathbf{V}$ and $\mathbf{WH}$, an iterative scheme with multiplicative update rules can be used as in [4]. In [5] it is proven that a point which is invariant under these update rules is also invariant under the update formulas of the algorithm proposed in [6] for probabilistic latent semantic analysis (PLSA). This means that a solution of NMF with divergence criterion is also a solution of the PLSA algorithm and vice versa. In fact, PLSA can also be interpreted as a non-negative factorisation of a non-negative observation matrix $\mathbf{V}$, using a divergence criterion. One can translate the PLSA algorithm into the following update rules:

$$\mathbf{W}_{ij}^{(t+1)} = \mathbf{W}_{ij}^{(t)} \sum_a \frac{\mathbf{H}_{ja}^{(t)} \mathbf{V}_{ia}}{(\mathbf{W}^{(t)}\mathbf{H}^{(t)})_{ia}} \tag{3a}$$

$$\mathbf{H}_{ij}^{(t+1)} = \mathbf{H}_{ij}^{(t)} \frac{\sum_a \frac{\mathbf{W}_{ai}^{(t)} \mathbf{V}_{aj}}{(\mathbf{W}^{(t)}\mathbf{H}^{(t)})_{aj}}}{\sum_a \mathbf{W}_{ai}^{(t+1)}} \tag{3b}$$

with $\mathbf{W}^{(0)}$ a random matrix and $\mathbf{H}^{(0)}$ a random matrix with rows summing to one. The update formulas (3) are very similar to the update formulas for NMF with divergence criterion in [4]. Experiments show that both algorithms converge at the same speed to the same solution when using the same initialisation. In the experiments



(a) Result of NMF

(b) MEL-frequency filter bank

Fig. 1. Comparison of the NMF basis vectors with the Davis & Mermelstein MEL-frequency filter bank

proposed in this paper, the update (3) is used because of the absence of a normalisation step in the update of $\mathbf{W}$. When $\mathbf{H}$ is properly scaled, the normalisation of $\mathbf{W}$ is implicit because the PLSA model uses probability matrices.

3. CONSTRUCTING THE DATA MATRIX

The columns of the data matrix $\mathbf{V}$ contain spectral slices of continuous speech recordings from the TIMIT database, sampled at 16000 Hz. The speech signals are pre-emphasised using a first-order high-pass filter $H(z) = 1 - \alpha z^{-1}$ with $\alpha = 0.95$. The resulting signals are decomposed into frames of 25 ms with 10 ms frame shift. Each frame is Hamming-windowed and zero-padded resulting in frames $\mathbf{x}_i$ with 512 samples where i goes from 1 to approximately 1.13×10^6. Let $\mathbf{X}_i$ be the first 257 points of the FFT of $\mathbf{x}_i$. The power spectra $|\mathbf{X}_i|^2$ are normalised to obtain unity energy and placed in the columns of $\mathbf{V}$. The normalisation prevents that high energetic phonemes dominate the factorisation, neglecting low energy phonemes like fricatives (see section 5). Because $\mathbf{V}$ contains speech power spectra, its dynamic range is high. This is why the divergence cost criterion is used to assess the approximation, and not the MSE criterion. With a MSE criterion, the highest peaks in the matrix $\mathbf{V}$ would dominate the entire factorisation.

4. FACTORISATION

The matrix $\mathbf{V}$ is factorised by NMF with $r = 24$. This is the classical number of MEL-features used in most speech recognition systems sampling at 16000 Hz. Increasing this number leads only to a minor decrease of recognition errors.

Fig. 1(a) shows the matrix $\mathbf{W}$ found by NMF. The columns are scaled to obtain a maximum value of 1 and are permuted for intelligibility. Repeating the experiment with other initialisations always reveals a similar result with minor translations of the bands in the frequency spectrum and sometimes higher side lobes. The basis vectors in the columns of $\mathbf{W}$ are sparse and can be interpreted as frequency bands. This is remarkable since no constraints were imposed on the shape of the basis vectors. The coefficients in $\mathbf{H}$ show how much each band is activated to reconstruct the according spectrum in $\mathbf{V}$.

Figure 1(b) shows the Davis & Mermelstein MEL-frequency filter bank which is often used to generate feature vectors for speech recognition [7]. There is a remarkable similarity between this filter bank and the basis vectors found by NMF, both having wider subbands at higher frequencies. The MEL-scale used for the design of the MEL-frequency filter bank is based upon the frequency analysis performed by the basilar membrane in the cochlea of the inner ear. It models the fact that frequency resolution is lower at high frequencies than it is at low frequencies, leading to wider spectral bands at high frequency.

The fact that a similar result is obtained by analysis of human perceptual hearing (MEL scale) and analysis of human speech as done in this experiment, shows that the human speech production system and the hearing system are well matched to each other. It seems that we produce speech in such a way that our hearing system is able to capture as much information as possible.

5. ENERGY DEPENDENCE OF THE USED CRITERION

Although the divergence criterion is better suited than MSE to analyse data with a high dynamic range, there is still a linear dependence on energy which may influence the decomposition. To see this, we analyse the divergence cost for reconstructing one element v by x. According to (2), the divergence is

$$Div(v \parallel x) = v \log \frac{v}{x} - v + x \tag{4}$$

Let $\delta = \frac{x-v}{v}$ be the relative reconstruction error. We can then rewrite (4) as

$$\begin{aligned} Div(v \parallel x) &= v \log \frac{v}{(1+\delta)v} + \delta v \\ &= (\delta - \log(1+\delta))v \end{aligned}$$

This shows that the same relative reconstruction error leads to a higher cost for a high value than for a low value. This means that the divergence criterion is biased, favouring parts of the spectrum with high mean energy. The pre-emphasis and the normalisation of the columns of $\mathbf{V}$ partly solves this problem. However, when observing Fig. 2 showing the mean energy in each row of $\mathbf{V}$, it is clear that the energy is not uniformly distributed over the whole spectrum. Because of the higher mean energy in the first quarter of the spectrum, NMF has the tendency to reconstruct the low-frequency parts of the spectrum with a higher resolution than the rest of the spectrum. This could explain why we obtain smaller subbands at the

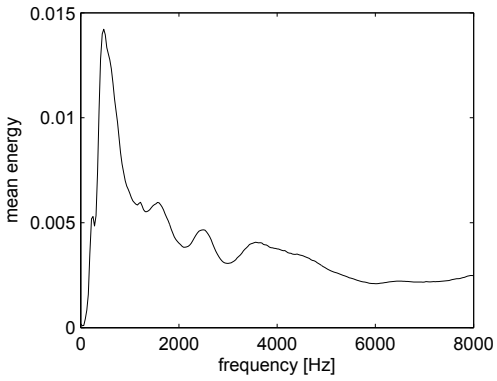


Fig. 2. Mean energy in $\mathbf{V}$ throughout the frequency spectrum

low frequencies. Therefore, we have to be careful concluding that NMF searches for basis vectors with most information content about the speech spectrum. It could be possible that we are measuring energy instead of information. In section 5.1, we will invalidate this experimentally, showing that NMF is not merely measuring energy. Furthermore, in section 5.2 we will argue that a larger focus on high energetic bands is necessary to obtain a good design.

5.1. Experimental counterarguments

When analysing Fig. 1(a), we see that the largest increase in bandwidth is found between 4000 and 8000 Hz. Notice that the second half of the spectrum is covered by only 20% of the basis vectors. However, when we observe Fig. 2, we see the steepest decrease in energy at about 1000 to 2000 Hz. If NMF was only measuring energy, the bandwidth would increase dramatically in this area and should be negligible in the area between 4000 and 8000 Hz, which is clearly not the case.

Furthermore, it is remarkable that the basis vectors found by NMF have contiguous FFT-bins without any gaps in the frequency response. Notice that we did not use any constraint to enforce this. This shows that NMF exploits the correlation between adjacent bins. Because formant bandwidth increases for subsequent formants, the correlation is spread over wider spectral areas at higher frequencies, leading to wider subbands. If perfect correlation would exist between FFT-bins, NMF would group according to correlation or information rather than energy. Therefore, the divergence criterion is useful, even though it may be biased.

5.2. Energy and information

Although energy and information are different concepts, they are intertwined: there can be no information without energy. It is necessary that bands with high mean energy have a stronger impact on the design: in the extreme case that there is no energy, NMF should not incorporate this in the design. The higher mean energy in the first part of the spectrum is due to a higher formant density, which results in a higher informative value. It is therefore desirable to place smaller and more bands in the spectral areas with high mean energy. If NMF would not have a focus on the spectral bands with high mean energy, this would be disadvantageous for the reconstruction of the low frequency areas where most structure and determinism can be found.

6. COMPARISON WITH HEARING MODELS

In this section, we compare our result with several hearing models found in literature. When comparing figures 1(a) and 1(b), an offset of approximately 200 Hz can be observed concerning the start of the first band. Experiments where the first band is forced to start at 0 Hz (this is possible by using the fact that zeros remain zero when using multiplicative update rules) always lead to a higher divergence. This shows that the offset is not caused by local optima. The actual cause is the lack of energy in the lower frequencies (cf. Fig. 2). This is due to pre-emphasis and the fact that the data in the TIMIT database is high-pass filtered, probably to get rid of a DC component. When testing on the Resource Management database [8], NMF does find a basis vector on these lower frequencies, but with an artifact on DC. The other basis vectors are very similar to the ones found with the TIMIT database.

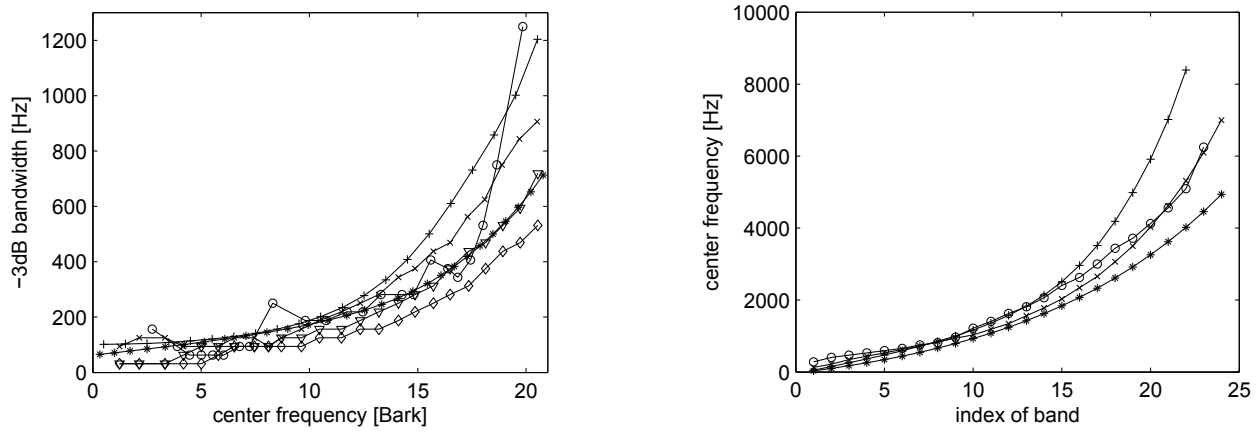
Fig. 3(a) shows the -3dB bandwidth vs. center frequency of each subband found with NMF (o), compared to other hearing models found in literature: the Davis & Mermelstein MEL-frequency filter bank ($\times$), the analytical ($\diamond$) and conventional ($\bigtriangledown$) gammatone filter bank[1] (AGT and CGT), subsequent non-overlapping rectangular bands of 100 MEL ($*$), and 1 Bark critical bands (+). The horizontal axis is mapped to the Bark scale for intelligibility. Fig. 3(b) shows the center frequencies of the NMF subbands, compared to the same hearing models. The AGT and CGT have the same center frequencies as the Davis & Mermelstein filter bank and are therefore omitted. The last band is omitted in both figures because the definition of the -3dB bandwidth and the center frequency of this last subband is ambiguous due to side effects. In both figures, the NMF curve has a similar shape as the other curves and lies -leaving some outliers out of consideration- within the same area. At 17 Bark or $\approx$4000 Hz, the slope of the curve representing the -3dB bandwidths of the NMF spectral bands becomes large when compared to the curves of the hearing models. This is not surprising, because this curve is the result of analysing speech signals. It is well-known that most of the information to understand speech is found below 4000 Hz.

7. RECOGNITION EXPERIMENTS

To compare the performance of the discovered feature set with the widely used MEL-features, a phoneme recognition experiment with triphone HMMs with a phone-level trigram is done on a test set different from the one used for the NMF factorisation. One might suggest that the columns of $\mathbf{H}$, containing the activation coefficients for each spectral band, could be used as feature vectors for speech recognition after performing a log-compression to lower the dynamic range of the coefficients. Although these coefficients do a good job in reconstructing the spectra, they are not well suited for use as feature vectors. The discovered features lead to phoneme error-rates (PER) of 31.7%, whereas using MEL-features results in a PER of 25.84%. The disappointing performance is caused by two problems.

First of all, NMF does not always preserve energy when reconstructing the spectra: sometimes it will omit or suppress a peak to reduce the divergence cost for reconstruction errors on neighbouring FFT-bins, leading to more recognition errors. This occurs mostly when a peak with small bandwidth appears between two adjacent NMF subbands. If NMF would reconstruct this peak, both subbands must be activated, resulting in high reconstruction errors on the neighbouring bins.

[1]The gammatone filter banks were generated with the MATLAB toolbox HUTear [9].



(a) -3dB bandwidths (b) center frequencies

Fig. 3. Comparison of the NMF basis vectors (o) with several hearing models: the Davis & Mermelstein MEL-frequency filter bank ($\times$), the analytical ($\diamond$) and conventional ($\bigtriangledown$) gammatone filter bank, subsequent non-overlapping rectangular bands of 100 MEL ($*$), and 1 Bark critical bands (+)

Another problem is the large amount of zero-coefficients in $\mathbf{H}$ (about 2.5%) due to overlap of the basis vectors. A basis vector can have a zero-coefficient if the energy in the corresponding frequency band is already present due to the activation of neigbouring bands. A $\log(x+\epsilon)$-compression maps the zero-coefficients to a single value $\log(\epsilon)$. This leads to artifacts in the cloud of data points in the feature space. These artifacts must be modelled by zero-variance Gaussian pdf's which are not handled well by the algorithm that estimates the parameters of the Gaussian mixtures. Eliminating these zeros by adding noise reduces the PER with approximately 3%.

Using the basis vectors as a filter bank -integrating the weighted spectral energy in each band- results in a feature set with approximately the same performance as MEL-features. The 24-dimensional feature vectors can be found in the columns of matrix $\mathbf{F}$ with $\mathbf{F} = \mathbf{W}^T\mathbf{V}$. Using these features gives a PER of 25.94%, which does not significantly differs from the MEL-based recognizer.

8. CONCLUSIONS

In this paper, we used NMF to factorise a data matrix with spectral slices of continuous speech to automatically find a feature set for speech recognition. The resulting decomposition has remarkable similarities to perceptually motivated filter bank designs such as the MEL-frequency filter bank. This supports the hypothesis that the biological systems for human hearing and speech production are well matched to each other, making human speech a well tuned and efficient communication system. Although we pointed out that the divergence criterion is biased, favouring the spectral areas with high mean energy, we argued that this does not affect the interpretation of our results: the decomposition exploits the correlation between adjacent FFT-bins, yielding basis vectors with large information content. Finally, we showed that when using the basis vectors as a filter bank to generate features, we obtain the same recognition performance as a MEL scaled filter bank with the same number of channels.

9. ACKNOWLEDGEMENT

This research was funded by 'Research Fund (Onderzoeksfonds) K.U.Leuven' (project no. OT/03/32/TBA), and by the European Commission under contract FP6-034362.

10. REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[2] P. Smaragdis, "Convolutive speech bases and their application to speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 1–12, January 2007.

[3] Michael S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, pp. 356–363, April 2002.

[4] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[5] E. Gaussier and C. Goutte, "Relation between plsa and nmf and implications," in *Proc. of the ACM SIGIR conference on research and development in information retrieval*, Salvador, Brazil, 2005, pp. 601–602.

[6] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of Uncertainty in Artificial Intelligence*, Stockholm, 1999.

[7] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 28, pp. 357–366, 1980.

[8] David S. Pallett, "Benchmark tests for darpa resource management database performance," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Glasgow, UK, May 1989, pp. 536–539.

[9] Aki Härmä and Kalle Palomäki, *HUTear - a Free Matlab Toolbox for Modeling of Human Auditory System*, www.acoustics.hut.fi/software/HUTear.