In Proceedings of International Joint Conference on Neural Networks (IJCNN'03), vol. 4, pp. 2758-2763 – Portland, OR, 2003.

Discovering hierarchical speech features using convolutional non-negative matrix factorization

Sven Behnke
International Computer Science Institute
1947 Center St., Berkeley, CA, 94704, USA
Email: behnke@icsi.berkeley.edu

Abstract— **Discovering a representation that reflects the structure of a dataset is a first step for many inference and learning methods. This paper aims at finding a hierarchy of localized speech features that can be interpreted as parts.**

Non-negative matrix factorization (NMF) has been proposed recently for the discovery of parts-based localized additive representations. Here, I propose a variant of this method, convolutional NMF, that enforces a particular local connectivity with shared weights. Analysis starts from a spectrogram. The hidden representations produced by convolutional NMF are input to the same analysis method at the next higher level.

Repeated application of convolutional NMF yields a sequence of increasingly abstract representations. These speech representations are parts-based, where complex higher-level parts are defined in terms of less complex lower-level ones.

I. Introduction

Unsupervised learning techniques [1] try to discover the structure underlying a dataset. Since data can always be interpreted in many different ways, some bias is needed to determine which aspects of the input's structure should be captured in the output of a learning machine. Principal component analysis (PCA), for instance, finds a low-dimensional representation of the data that captures most of its variance. On the other hand, slow feature analysis (SFA) [2] tries to minimize the variance of features as the signal undergoes a transformation, and independent component analysis (ICA) [3] seeks to explain the data as a mixture of independent sources.

In general, unsupervised learning has the goal of finding useful representations of the given data. Examples include: grouping examples to clusters, reduction of data dimensionality, discovering hidden causes of the data, or modeling the data density. If unsupervised learning is successful, the produced representations can be used for tasks such as data compression, outlier detection, and classification. It can also make other learning tasks, like pattern recognition, easier.

One of the most important problems in pattern recognition is the extraction of meaningful features from input signals. To compute symbolic information, such as the class of an observed object, it is often useful to aggregate characteristic aspects of the observation into a feature vector that is presented to a classification system. This generally reduces the dimensionality of the data and facilitates generalization by discarding aspects of the signal that correspond to variances not relevant for classification or to noise.

A variety of feature extraction methods exist, e.g. for the problem of automatic speech recognition (ASR). Most of them are based on the spectral decomposition of a short time frame (typically 10ms) of speech. Mel-frequency cepstral coefficients (MFCC) [4] and perceptual linear predictive coefficients (PLP) [5] form the basis of many ASR systems. A drawback of the spectral features is that they are quite sensitive to changes in the environment such as channel variations or noise. Consequently, the performance of recognizers based on spectral features rapidly degrades in realistic communication environments. Psychoacoustic studies [6] and recent neurophysiological experiments [7] suggest that the peripheral auditory system in humans integrates information from much larger time spans. Spectro-temporal receptive fields that cover several hundred milliseconds are a linear approximation to the non-linear responses of neurons in the auditory cortex.

This paper proposes a hierarchical approach to speech feature extraction. Complex features are defined in terms of less complex ones. This is motivated by the hierarchical structure that speech exhibits. Different speech components exist in different time scales. The periodic glottal excitation in voiced speech has a timescale of few milliseconds and leads to the perception of pitch. Phonemes are the atoms of speech, having a temporal extent of tens of milliseconds. Syllables are composed of several phonemes, extending over hundreds of milliseconds. Words are based on syllables and span even longer time intervals and phrases that can extend over several seconds consist of multiple words. An adequate model of speech must represent the components at multiple time-scales in order to capture the correlations at different levels of this hierarchy and between adjacent levels.

For images, such hierarchical models have been discovered by unsupervised learning techniques and successfully used for object recognition tasks in the Neural Abstraction Pyramid architecture [8], [9]. This architecture has also been used for image reconstruction [10] and face localization [11]. Figure 1 illustrates its main properties. The input signal is represented at different levels of abstraction. As the spatial resolution of these representations decreases, feature diversity and invariance to transformations increase. The representations are produced by simple processing elements that interact locally. Horizontal and vertical feedback loops allow for consideration of contextual information to iteratively resolve local ambiguities.

Lee and Seung [12] recently proposed a generative data model, called non-negative matrix factorization (NMF). It can be interpreted in terms of vertical feedback. Here, I propose to use NMF in the Neural Abstraction Pyramid architecture to

discover hierarchical speech features.

The paper is organized as follows. The next section reviews the basic NMF technique and highlights some of its properties. Section III details the extensions for convolutional NMF and its repeated application in a hierarchy. The dataset used for the experiments is described in Section IV, and Section V illustrates the discovered hierarchical speech features.

II. Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF) [12] decomposes a $n \times m$ non-negative matrix V approximately into non-negative matrix factors: $V \approx WH$. The m columns of V consist of n-dimensional data vectors. The r columns of W contain basis vectors of dimension n. Each data vector is represented by a column of H that contains r coefficients. This corresponds to a two-layered neural network which represents the data vector in a visible layer and the coefficients in a hidden layer. The matrix W describes the weights that connect both layers.

One measure of the factorization quality is the square of the Euclidean distance $\|A-B\|^2 = \sum_{ij}(A_{ij}-B_{ij})^2$ between V and its reconstruction WH. $\|V-WH\|^2$ is minimized by:

$$H_{a\mu} \leftarrow H_{a\mu}\frac{(W^TV)_{a\mu}}{(W^TWH)_{a\mu}}; \qquad W_{ia} \leftarrow W_{ia}\frac{(VH^T)_{ia}}{(WHH^T)_{ia}}.$$

Another quality measure is the divergence

$$D(A\|B) = \sum_{ij}(A_{ij}\log\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}). \qquad (1)$$

$D(V\|WH)$ is minimized by:

$$H_{a\mu} \leftarrow H_{a\mu}\frac{\sum_i W_{ia}V_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}}; \qquad (2)$$

$$W_{ia} \leftarrow W_{ia}\frac{\sum_\mu H_{a\mu}V_{i\mu}/(WH)_{i\mu}}{\sum_\nu H_{a\nu}}. \qquad (3)$$

Lee and Seung [13] proved that these update rules find local minima of the respective objective functions. Each update consists of a multiplicative factor that is unity if $V = WH$. The multiplicative update does not change the sign of W or H. Hence, if they are initialized to positive values no further constraints are necessary to enforce their non-negativity.

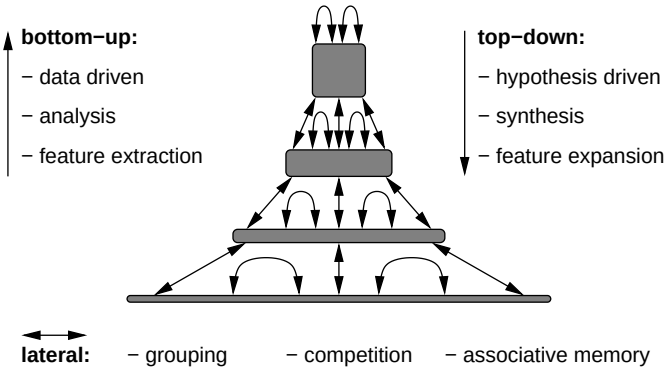

Fig. 1. Integration of bottom-up, lateral, and top-down processing in the Neural Abstraction Pyramid. Signals are represented at different levels of abstraction. As the spatial resolution decreases, feature diversity and invariance to transformations increase. Local recurrent connections mediate interactions of simple processing elements.

The model was applied to a dataset of normalized facial images. The basis vectors found by minimizing the divergence consist of localized patches of high values that resemble typical dark face regions, such as the eyes and the shadow of the nose. They can be interpreted as parts. Both, the weights and the coefficients of the parts-based encoding h, contained a large number of vanishing components. The reason for this is that the model is only allowed to add positively weighted non-negative basis-vectors to the reconstruction. Thus, different contributions do not cancel out, as for instance in PCA.

Although the generative model is linear, inference of the hidden representation h from an image v is highly non-linear. The reason for this is the non-negativity constraint. It is not clear how the best hidden representation could be computed directly from W and v. However, as seen above, h can be computed by a simple iterative scheme. Because learning of weights should occur on a much slower time-scale than this inference, W can be regarded as constant. Then only the update-equations for H remain.

When minimizing $\|v-Wh\|^2$, h is sent in the top-down direction through W. Wh has dimension n and is passed in bottom-up direction through W^T. The resulting vector W^TWh has the same number r of components as h. It is compared to W^Tv, which is the image v passed in bottom-up direction through W^T. The comparison is done by element-wise division yielding a vector of ones if the reconstruction is perfect: $v = Wh$. In this case, h is not changed.

When minimizing $D(v\|Wh)$, the similarity of v and its top-down reconstruction Wh is measured in the bottom-layer of the network by element-wise division $v_i/(Wh)_i$. The n-dimensional similarity-vector is passed in bottom-up direction through W^T, yielding a vector of dimension r. Its components are scaled down with the element-wise inverse of the vector of ones passed through W^T to make the update factors for h unity, if the reconstruction is perfect.

This scheme of expanding the hidden representation to the visible layer, measuring differences to the observations in the visible layer, contracting the deviations to the hidden layer, and updating the estimate resembles the operation of a Kalman filter [14]. The difference is that in a Kalman filter deviations are measured as differences and update is additive, while in the non-negative matrix factorization deviations are measured with quotients and updates are multiplicative. Because the optimized function is convex for a fixed W, the iterative algorithm is guaranteed to find the optimal solution.

III. Convolutional NMF in a Hierarchy

A. Convolutional NMF

If data vector v has many elements and the number of hidden features in h is high as well, the matrix W is large. Moreover, if the analyzed data vector is an image or another two-dimensional signal, NMF produces localized features, since neighboring data elements tend to be more correlated than far-apart ones. In this case it is natural to enforce a

sparseness structure of W from the beginning that only allows for local feature extraction.

The particular sparseness structure used here resembles the local connectivity in the Neural Abstraction Pyramid architecture [9]. The visible and the hidden layer representing v and h, respectively, are partitioned into two-dimensional feature arrays. The spectral and temporal resolutions of these arrays in the hidden layer is half the one of the visible layer. Thus, a hidden feature cell at position (i, j) corresponds to the position $(2i, 2j)$ in the visible layer. In general, the number of hidden feature arrays is larger than the number of visible ones. Here, they differ by a factor of two. Hidden feature cells can access only the corresponding patch in the visible layer. Here, I use overlapping windows of size 4×4. All elements of W outside this window are initialized to zero and never become nonzero, since the weight update is multiplicative. The NMF algorithm still can decide which cells of the patch are to be combined to form a part.

In the Neural Abstraction Pyramid, the same weights are used at all positions of a feature array. Here, this corresponds to the assumption that the same basic parts can be used to encode speech at different points in time and at different parts of the frequency range. This weight sharing reduces the number of free parameters and allows for the sharing of examples. Hence, good generalization can be expected even from a small training set. The sparseness structure and the weight sharing can be exploited when computing equations (2) and (3) for inference and learning. For instance, $v = Wh$ can be computed by super-sampling the feature arrays of h, convolution with a 4×4 kernel, and element-wise summation of the results. If the kernel size is small compared to the array size, this yields a large reduction in computational cost, as compared to a fully occupied matrix W.

B. Recursive Application

Unlike many other methods of unsupervised learning, convolutional NMF is suitable for recursive application by using its hidden representation as input to the next stage of analysis. This is possible, since the statistics of hidden layer coefficients are similar to the ones in the visible layer. In both layers activations are non-negative and sparse.

The recursive application of convolutional NMF uses a parts-based encoding as input to an analysis that produces an encoding based on larger parts. This yields a sequence of increasingly abstract representations, where the spatial resolution decreases and feature diversity increases.

IV. Dataset and Preprocessing

To test the proposed feature extraction method, I used speech data from the ICSI meeting corpus [15]. The part used here consists of read continuous digit sequences recorded from a close talking microphone. The speech signal has been downsampled to 8kHz and a pre-emphasis factor of 0.97 has been applied to attenuate lower frequencies.

I used a 975ms Hanning window with step size 487.5ms to produce an input signal for feature analysis of constant size. This signal is decomposed using short-time Fourier Transformation (STFT) computed for 25ms Hanning windows with a step size of 10ms. After taking the magnitude the 96×96 spectrogram represents the energy of 96 time steps and 96 frequency bands linearly spaced from 0 to 3.840kHz.

V. Experimental Results

Convolutional NMF is applied five times to one hundred spectrograms, until the representation consists of 32 features of size 3×3. For each level, the nonzero elements of the weight matrix W_i and the hidden representation H_i are initialized to random values uniformly distributed in the interval $[0.1, 0.2]$. The NMF equations (2) and (3) minimizing the divergence (1) are then iterated 200 times. Figure 2 displays the weights determined by this procedure. One can observe that the weight matrix is sparsely populated and that every visible feature contributes to some hidden feature and vice versa.

Since the expansion of hidden representation is linear, one can get a more interpretable impression of the hidden representation by setting individual features to one and expanding them all the way down to the spectrogram. This is done in Fig. 3 for the first three levels of the hierarchy and in Fig. 4 for the remaining two levels. As one can observe, the features in the higher levels get more and more diverse. They are all localized and can be interpreted as parts. Many features look like vertical lines. They are localized precisely in time, but cover a larger range of frequencies. Other features look more like blobs, with localization in time as well as in frequency with medium precision. Some features in the upper levels have more than one active region. They look like two adjacent vertical lines or two blobs that are displaced by the typical formant distance of about 1kHz. In Fig. 5 the expansions of all hidden features of a single column have been added. One can observe that the features are localized at different positions and cover the column's area of the time-frequency plane.

Since the responses of the hidden feature detectors depend non-linearly on the input, the linear expansions characterize

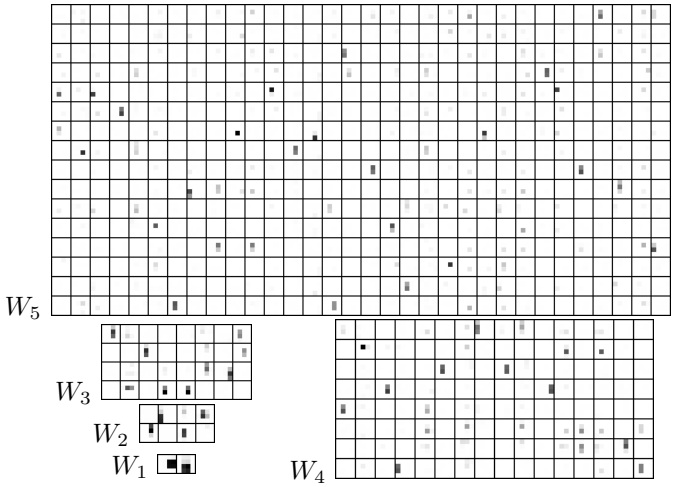

Fig. 2. Weights discovered by recursive application of convolutional NMF. On each level, the columns correspond to the hidden features and the rows to the visible ones. Normalization is such that the strongest weight is black.

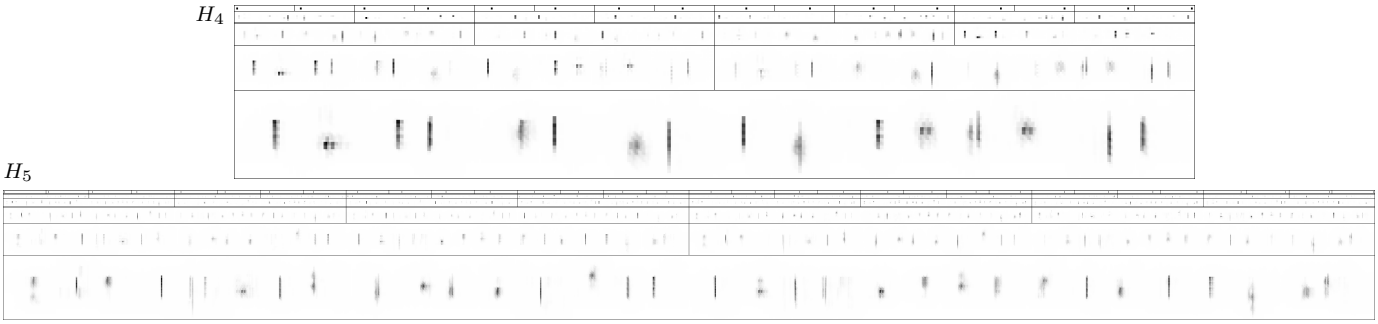

Fig. 4. Expanding hidden features. The remaining two levels of the hierarchy are shown. Refer to Fig. 3 for details.

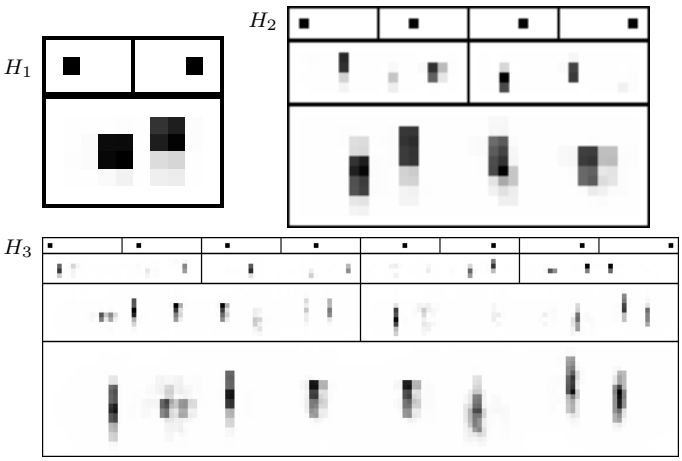

Fig. 3. Expanding hidden features. The first three levels of the hierarchy are shown. For the different hidden features one feature cell is set to one while all others are zero. This produces all used parts in the bottom rows.

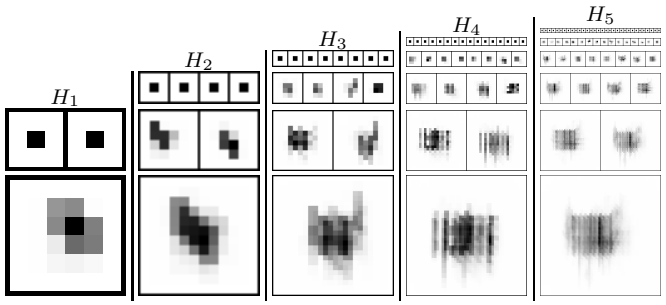

Fig. 5. Sum of the expanded hidden features of the central column.

their behavior only partially. Another possibility is to look at the stimuli from the training set that caused their highest responses. This is done in Fig. 6. One can observe that the features are active in quite different contexts and that the stimuli exciting a feature vary in many aspects. The parts-based representations abstract from these variations by discarding some details.

Which aspects of the speech signal are preserved when going from one layer of the feature hierarchy to the next can be observed in Fig. 7. It shows the hierarchical feature extraction for three speech signals along with the expansion of the abstract representations. The horizontal lines that correspond to the harmonics of the fundamental frequency are the first significant details which cannot be reconstructed. In reconstructions from the most abstract features one can clearly see that the parts-based representation has a much better resolution in time than in frequency. While the precise timing of plosives is reconstructed, only a coarse approximation of the formant structure is produced. This preference for time-centered energy over frequency centered one is most likely caused by the high weight of reconstruction errors for energy-rich plosives and onsets in the minimized divergence (1). Futhermore, one can observe that the hidden representations are indeed sparse and that all features participate in the encoding.

VI. Conclusion and Future Work

The repeated application of convolutional NMF discovered a hierarchy of meaningful speech features. Since the reconstructed signal is a sum of such features, they can be interpreted as parts. The weight matrices describing these parts are sparse, as are the part-based distributed representations.

Sparse weight matrices allow for pruning without much loss. Sparse codes combine advantages of local and dense codes while avoiding most of their drawbacks [16]. Codes with small activity ratios can still have sufficiently high representational capacity, while the number of input-output pairs that can be stored in an associative memory is far greater for sparse than for dense patterns. The speed of learning with sparse patterns is fast and generalization to partially matching patterns can be expected. Finally, multiple objects can be represented simultaneously in the same feature vector, without much interference.

The weight sharing used in the proposed approach leads to the reuse of parts at different locations of the input signal. It limits the number of free parameters and makes it possible to extract multiple examples from one input signal. Both effects improve generalization.

By construction it is clear that the abstract representation at layer l is translated by k columns when the layer 0 input signal is translated by 2^k columns. However, translation of the input signal by intermediate distances will not result in a mere translation of the abstract representations by sub-column distances. Consider, for example, the family of translated vertical features that are highly localized in time. When shifting the input signal slightly in time the active feature cell will become inactive and the feature cell representing the neighboring vertical stripe is activated. This continues,

until the translation corresponds to the full column distance. Then the original feature becomes active again, but at the neighboring column. Hence, if a translation-invariant detector of vertical energy would be desired, pooling of the outputs of all vertical feature detectors of a column would be needed. Such pooling could be achieved e.g. by summation or by computing the maximal response.

So far, neighboring feature cells interact only through their overlapping projections to the lower layer. It seems promising to investigate the effects of adding direct lateral interactions. They could be used to implement mechanisms, such as grouping, competition, and associative memory.

ACKNOWLEDGMENT

This work was supported by a fellowship within the postdoc program of the German Academic Exchange Service (DAAD).

REFERENCES

[1] H. B. Barlow, "Unsupervised learning," *Neural Computation*, vol. 1, no. 3, pp. 295–311, 1989.

[2] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Computation*, vol. 14, no. 4, pp. 715–770, 2002.

[3] C. Jutten and J. Herault, "Blind separation of sources. an adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–31, 1991.

[4] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[6] J. L. Verhey, "Psychoacoustics of spectro-temporal effects in masking and loudness perception," Universität Oldenburg," Dissertation Thesis, 1999.

[7] C. K. Machens, M. Wehr, and A. M. Zador, "Spectro-temporal receptive fields of subthreshold responses in auditory cortex," in *Proceedings of NIPS 2002: Advances in Neural Information Processing Systems 15*, to appear 2003.

[8] S. Behnke, "Hebbian learning and competition in the Neural Abstraction Pyramid," in *Proceedings of IJCNN'99 – Washington, DC, paper #491*, 1999.

[9] ——, "Hierarchical neural networks for image interpretation," Freie Universität Berlin," Dissertation Thesis, 2002.

[10] ——, "Learning iterative image reconstruction in the Neural Abstraction Pyramid," *International Journal of Computational Intelligence and Applications, Special Issue on Neural Networks at IJCAI-2001*, vol. 1, no. 4, pp. 427–438, 2001.

[11] ——, "Learning face localization using hierarchical recurrent networks," in *Proceedings of International Conference on Artificial Neural Networks: ICANN 2002 – Madrid*, ser. LNCS, vol. 2415. Springer, 2002, pp. 1319–1324.

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

[13] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds. MIT Press, 2001, pp. 556–562.

[14] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[15] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003 – Hong Kong)*, to appear April 2003.

[16] P. Földiak, "Sparse coding in the primate cortex," in *The Handbook of Brain Theory and Neural Networks, Second edition*, M. A. Arbib, Ed. MIT Press, 2002.

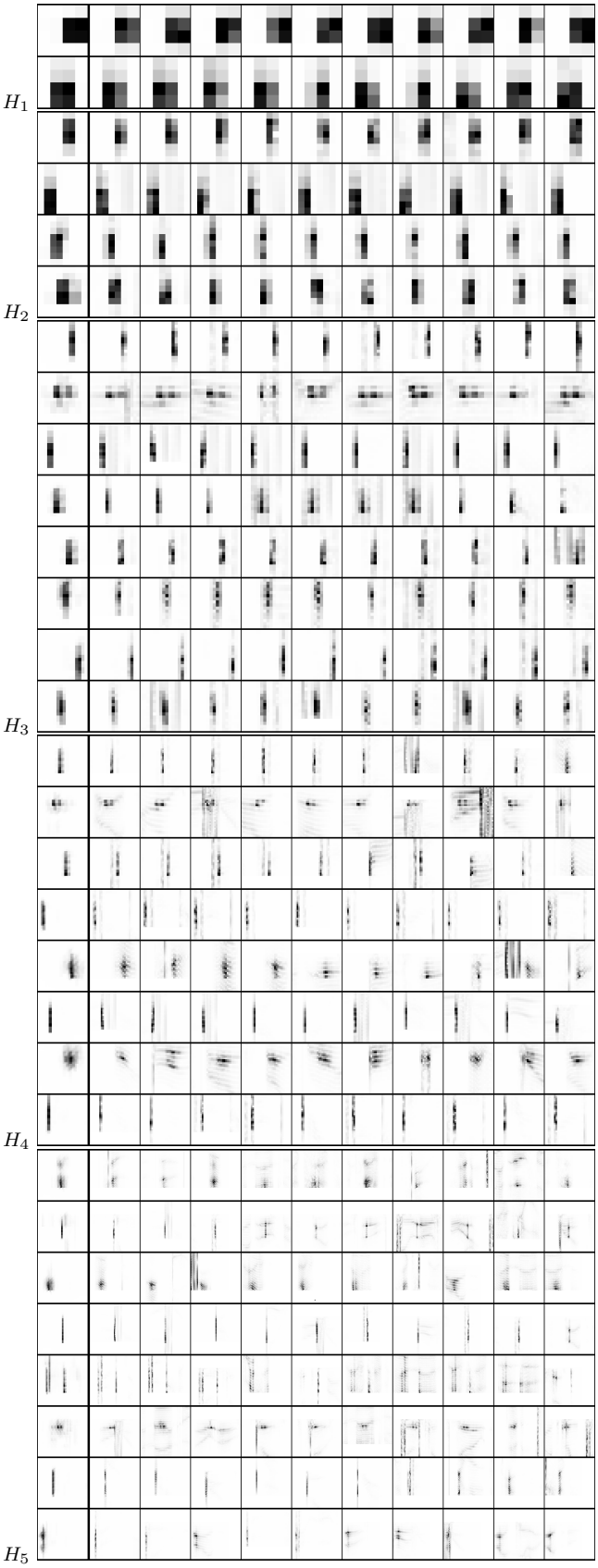

Fig. 6. Best stimuli. The leftmost column shows the expanded hidden features. Each row shows the ten best stimuli for a feature. The contrast of the spectrogram is lowered to one tenth outside the active region of the feature. For the two topmost levels 4 and 5, only eight features are shown.

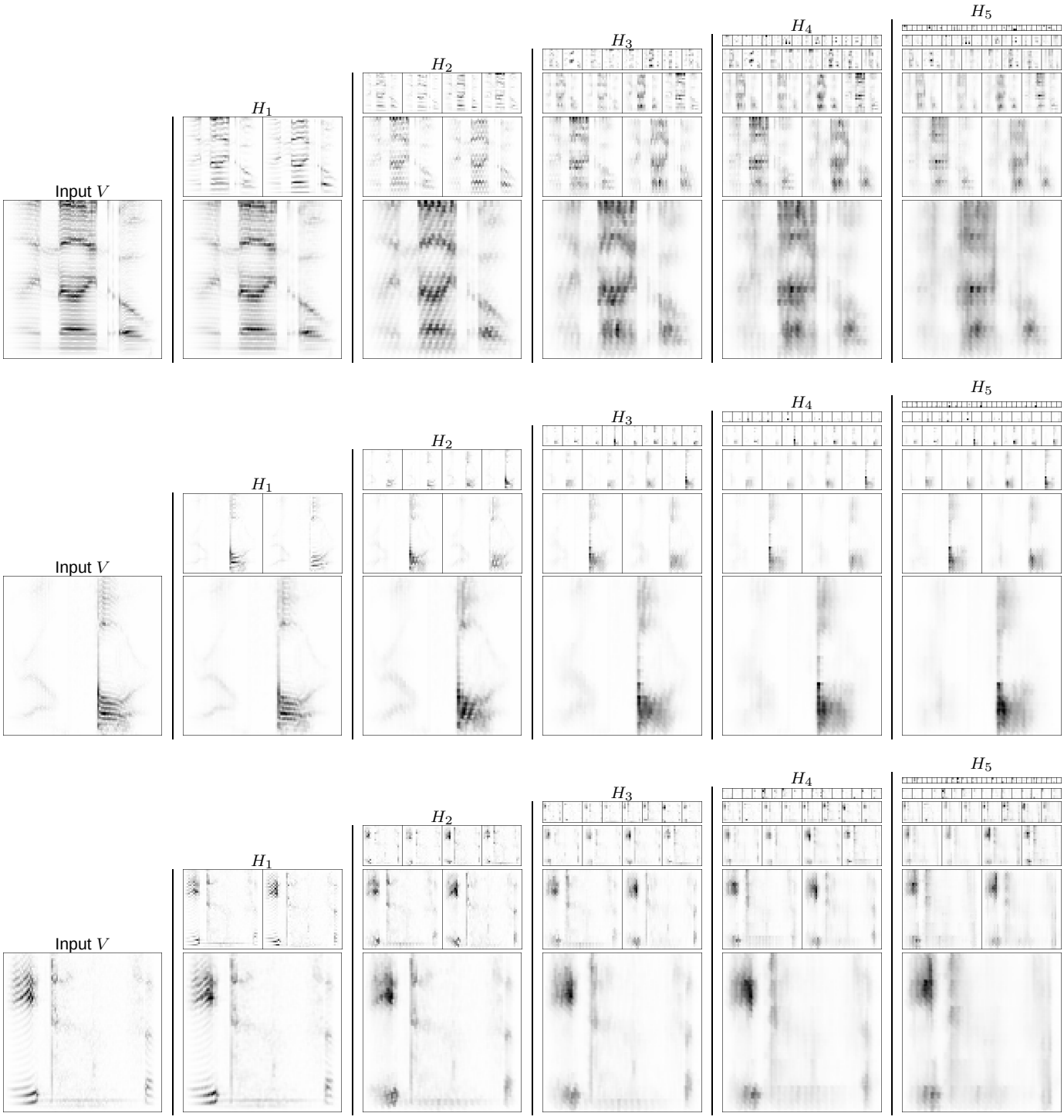

Fig. 7. Hierarchical speech features. The leftmost column shows the preprocessed spectrogram used as input for the analysis. The remaining columns show the successive construction of hidden representations and the expansion of these. As the size of the representation shrinks from $1\times96\times96$ to $32\times3\times3$, more and more details of the spectrogram are discarded. Details on the time-axis are better preserved than frequency details.