

# **Word Recognition with a Hierarchical Neural Network**

**Xavier Domont, Martin Heckmann, Heiko Wersing,  
Frank Joublin, Stefan Menzel, Bernhard Sendhoff,  
Christian Goerick**

**2007**

**Preprint:**

This is an accepted article published in Advances in Nonlinear Speech Processing. The final authenticated version is available online at:  
[https://doi.org/\[DOI not available\]](https://doi.org/[DOI not available])

# Word Recognition with a Hierarchical Neural Network

Xavier Domont<sup>1,2</sup>, Martin Heckmann<sup>1</sup>, Heiko Wersing<sup>1</sup>, Frank Joublin<sup>1</sup>,  
Stefan Menzel<sup>1</sup>, Bernhard Sendhoff<sup>1</sup>, and Christian Goerick<sup>1</sup>

<sup>1</sup> Honda Research Institute Europe GmbH,  
D-63073 Offenbach am Main, Germany  
[martin.heckmann@honda-ri.de](mailto:martin.heckmann@honda-ri.de)

<sup>2</sup> Technische Universität Darmstadt, Control Theory and Robotics Lab,  
D-64283 Darmstadt, Germany  
[xavier.domont@rtr.tu-darmstadt.de](mailto:xavier.domont@rtr.tu-darmstadt.de)

**Abstract.** In this paper we propose a feedforward neural network for syllable recognition. The core of the recognition system is based on a hierarchical architecture initially developed for visual object recognition. We show that, given the similarities between the primary auditory and visual cortexes, such a system can successfully be used for speech recognition. Syllables are used as basic units for the recognition. Their spectrograms, computed using a Gammatone filterbank, are interpreted as images and subsequently feed into the neural network after a preprocessing step that enhances the formant frequencies and normalizes the length of the syllables. The performance of our system has been analyzed on the recognition of 25 different monosyllabic words. The parameters of the architecture have been optimized using an evolutionary strategy. Compared to the Sphinx-4 speech recognition system, our system achieves better robustness and generalization capabilities in noisy conditions.

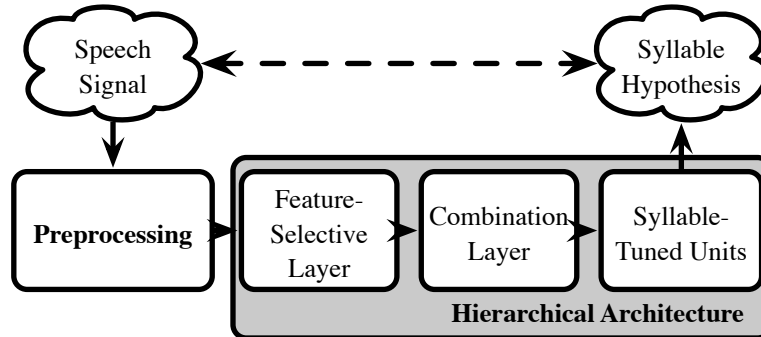
**Keywords:** speech recognition, robust features, feed-forward architecture.

## 1 Introduction

Conventional speech recognition systems perform very well in clean scenarios but their performance drastically decreases in noisy environments. This poor performance in adverse conditions prohibits the application of such systems for many scenarios, especially our target scenario, the control of a humanoid robot. In contrast to this, human speech perception is far less susceptible to such distortions [1].

In this article we present a speech recognition system with a higher robustness towards noise and reverberation. This system is based on a feedforward neural network inspired from an object recognition system.

Several studies have shown that auditory and visual primary cortices show substantial similarities. In 1988 Sur et al. have shown that the primary auditory cortex of young ferrets is plastic enough to allow the ferrets to attain visual



**Fig. 1.** Overview of the word recognition system

perception via the auditory cortex [2]. More recently Shamma determined the shape of the time-frequency receptive fields in the primary auditory cortex of newborn ferrets [3]. They are selective to modulations in the time-frequency domain and, as in the visual cortex, have Gabor-like shapes. These receptive fields have been modeled by Chin [4] and used for source separation [5] and speech detection [6].

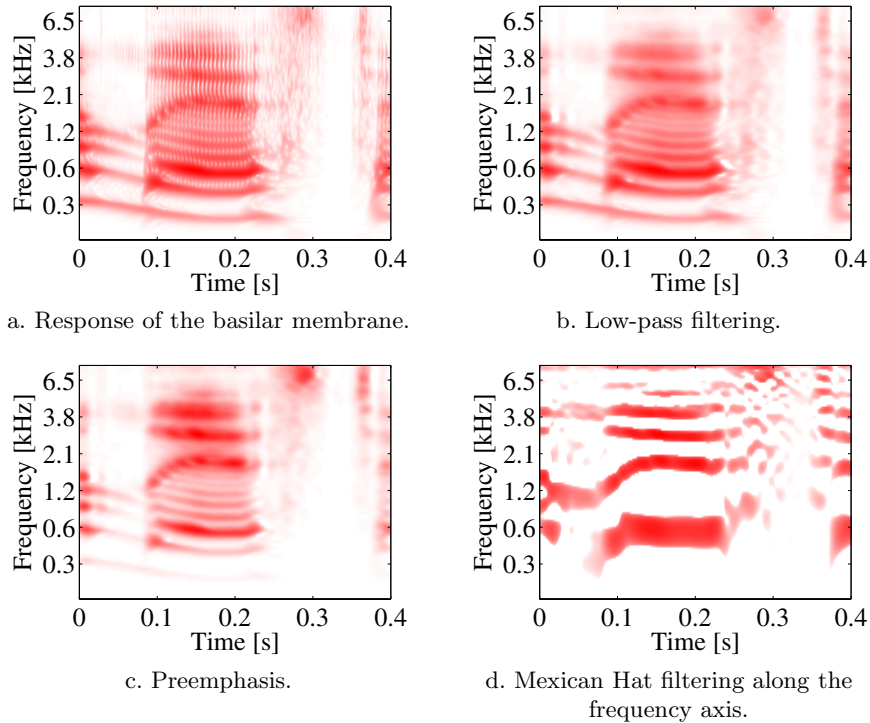
As Gabor-like filters are extensively used in object recognition systems [7,8], we decided to develop a system for speech recognition by adapting the feed-forward neural network initially developed by Wersing and Körner for object recognition [8]. This approach is similar to the spectro-temporal features and the direct recognition on spectrograms proposed by Kleinschmidt in [9].

Syllables are the basic units for speech production and show less coarticulatory effects across their boundaries. Therefore, we believe that they are the adequate speech units for our biologically-inspired system. Moreover, the syllable segmentation required for the training of the system seems biologically plausible for speech acquisition.

The building blocks of the system (Fig. 1) are detailed in the following sections. After explaining how we optimized the parameters of the architecture using an evolutionary strategy, we will compare our results to a state of the art speech recognition system and conclude with a discussion of the obtained results.

## 2 Preprocessing of the Spectrogram

The preprocessing mainly aims at transforming a previously segmented speech signal, corresponding to one syllable, into an "image" that is fed into the hierarchical recognition architecture. A two-dimensional representation of a signal is obtained by computing its spectrogram. In addition to the phonetic information, the speech signal also contains many speaker and recording specific information. As the phonetic information is chiefly conveyed by the formant trajectories, we enhance them in the spectrograms prior to recognition.



**Fig. 2.** Overview of the preprocessing step for the word "list" spoken by a female American speaker. The 128 channels logarithmically span the frequency range from 80 Hz to 8 kHz. The harmonic structure has been removed using a filtering along the frequency axis.

We used a Gammatone filterbank to compute the spectrogram of the signal. It models the response of the basilar membrane in the human inner ear and is, therefore, adapted to a biology-inspired system. The signal's sampling frequency is 16 kHz. The filterbank has 128 channels ranging from 80 Hz to 8 kHz and follows the implementation given in [10]. Figure 2 shows the response of the Gammatone filterbank after rectification (a.) and low-pass filtering (b.). To compensate for the influence of the speech excitation signal, the high frequencies are emphasized by +6 dB per octave resulting in a flattened spectrogram (Fig. 2 c.). Next, the formant frequencies are enhanced by filtering along the channel axis using mexican-hat filters (Fig. 2 d.), only the positive values are kept. For the filtering the size of the kernel is channel-dependent, varying from 90 Hz for low frequencies to 120 Hz for high frequencies. This takes the logarithmic arrangement of the center frequencies in the Gammatone filterbank into account.

Finally, the length of the spectrogram is scaled using linear interpolation so that all the spectrograms feeding the recognition hierarchy have the same size. The sampling rate is then reduced to 100 Hz. By doing so syllables of different

lengths are scaled to the same length. This relies on the assumption that a linear scaling can handle variations in the length of the same syllable uttered at different speaking rates. However, these are known to be non-linear. In particular, some parts of the signal, like vowels, are more affected by variation in the speech rate than other parts, e.g. plosives. The generalization over these variations is a main challenge in the recognition task. In order to also assess the performance of the recognition hierarchy independent of this non-linear scaling, we applied the Dynamic Time Warping (DTW) method to the spectrograms. For each syllable, we selected one single repetition as reference template and aligned the other by DTW.

Afterwards the syllables were again scaled to the same length and downsampled. At the output of the preprocessing stage the spectrograms feeding the recognition hierarchy have all the size of  $128 \times 128$ , i.e. 128 time frames over 128 frequency channels. Note, however, that the application of DTW requires that a hypothesis for the syllable is available. Thus, it cannot easily be applied to a real recognition test.

### 3 The Recognition Hierarchy

The preprocessed two-dimensional spectrogram is from now on considered to be an image and feeds into a feedforward architecture initially aimed at visual object recognition. However, the structure of spectrograms differs from the structure of images taken from objects and, while keeping the overall layout of the network described in [8], the receptive fields and the parameters of the neurons were retrained for the task of syllable recognition. The recognition hierarchy is illustrated in Fig. 3.

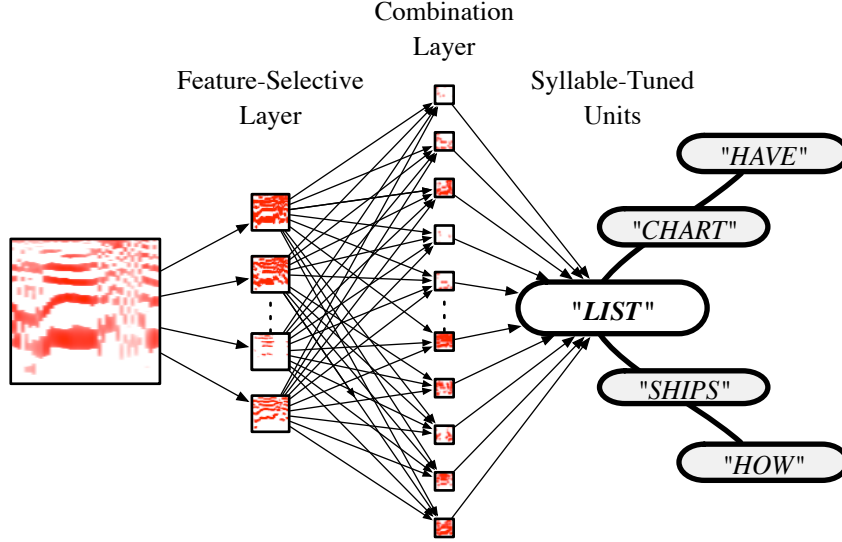
#### 3.1 Feature-Selective Layer

The first feature-matching stage consists of a linear receptive field summation, a Winner-Take-Most (WTM) and a pooling mechanism. The preprocessed spectrogram is first filtered by eight different Gabor-like filters. The purpose of these filters is to extract local features from the spectrogram. In [8] the receptive fields were chosen as four first-order even Gabor filters. For syllable recognition, 8 receptive fields were learned using independent component analysis on 3500 randomly selected local patches of preprocessed spectrograms.

The WTM competition mechanism between features at the same position introduces nonlinearity into the system. The value  $r_l(t, f)$  of the spectrogram in the  $l$ th neuron of the feature-selective layer after the WTM competition is given at the position  $(t, f)$  by the following equation:

$$r_l(t, f) = \begin{cases} 0, & \text{if } \frac{q_l(t, f)}{M(t, f)} < \gamma_1 \text{ or } M(t, f) = 0 \\ \frac{q_l(t, f) - \gamma_1 M(t, f)}{1 - \gamma_1}, & \text{else} \end{cases} \quad (1)$$

where  $q_l(t, f)$  is the value of the spectrogram before the WTM competition,  $M(t, f) = \max_k q_k(t, f)$  the maximal value at position  $(t, f)$  over the eight neurons and  $0 \leq \gamma_1 \leq 1$  is a parameter controlling the strength of the competition.



**Fig. 3.** The system is based on a feedforward architecture with weight-sharing and a succession of feature sensitive matching and pooling stages. It comprises three stages arranged in a processing hierarchy.

A threshold  $\theta_1$  is applied to the activity  $r_l(t, f)$ . This threshold is common for all the neurons in the layer. The pooling performs a downsampling of the spectrogram by four in both time and frequency direction. It is done by a Gaussian receptive field with width  $\sigma_1$ . The feature-selective layer transforms the  $128 \times 128$  original spectrogram to eight  $32 \times 32$  spectrogram feature maps.

### 3.2 Combination Layer

The goal of the combination layer is to detect relevant local feature combinations in the first layer. Similar to the previous layer it consists of a linear receptive field summation, a Winner-Take-Most and a pooling mechanism. These combination cells are learned using the non-negative sparse coding method (NNSC) as in [8], however no invariance transformations have been implemented at this stage. Similarly to Non-Negative Matrix Factorization (NMF), the NNSC method decomposes data vectors  $\mathbf{I}^p$  into linear combinations (with non-negative weights  $s_i^p$ ) of non-negative features  $\mathbf{w}_i$  by minimizing the following cost function:

$$E = \sum_p \|\mathbf{I}^p - \sum_i s_i^p \mathbf{w}_i\|^2 + \beta \sum_p \sum_i |s_i^p|.$$

NNSC differs from NMF by the presence of a sparsity enforcing term in the cost function, controlled by the parameter  $\beta$ , which aims at limiting the number of non-zero coefficients required for the reconstruction. Consequently, if a feature appears often in the data, it will be learned, even if it can be obtained by a combination of two or more other features. Therefore, the NNSC is expected

to learn complex and global features appearing in the data. An comprehensive description of this method can be found in [11].

For the proposed syllable recognition system 50 complex features  $\mathbf{w}_i$  have been learned from image patches extracted from the output of the feature-selective layer. At last, a WTM competition ( $\gamma_2, \theta_2$ ) and pooling ( $\sigma_2$ ) are applied to the 50 neurons and their size is reduced to  $16 \times 16$ .

### 3.3 Syllable-Tuned Units

In the last stage of the architecture, linear discriminant classifiers are learned based on the output of the combination layer. A classical gradient descent is used for this supervised learning including an early stopping mechanism to avoid overfitting. The obtained classifiers are called Syllable-Tuned Units (STUs) in reference to the View-Tuned Units used in [7] and [8].

Due to the high dimensionality (640) and sparseness of the features after the combination layer learning the STUs is unproblematic.

## 4 Optimization of the Architecture

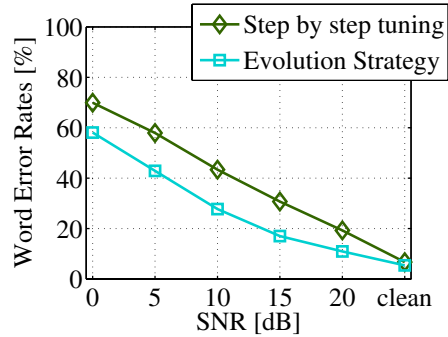
The performance of the recognition highly depends on the choice of the non-linearities present in the hidden layers of the architecture, i.e. the coefficients and the thresholds of the WTM competitions (Eq. 1) and the width of the poolings. The six parameters ( $\gamma_{1,2}, \theta_{1,2}$  and  $\sigma_{1,2}$ ) have to be tuned simultaneously and the receptive field of the combination layer as well as the Syllable-Tuned Units have to be learned at each iteration, similarly to the method used in [12].

Practically, this tuning of the model parameter set has been realized within an evolutionary optimization aiming at maximizing the recognition performance in a clean speech scenario. Due to the stochastic components and the use of a population of solutions evolutionary algorithms need more quality evaluations than other algorithms, but on the other hand they allow for a global search and are able to overcome local optima. In the present context, an evolutionary strategy with global step size adaptation (GSA-ES) has been applied relying on similar ranges of the object variables. Initially, standard values, see [13,14], have been used and then tuned in some test experiments to this specific task. Based on these experiments we have chosen a population size of 32 individuals. Each generation, the two individuals with the best performance have been chosen as parents for the next generation. The optimization parameters have been scaled and the initial global step size was set to 0.003.

Although the evolutionary optimization used a clean scenario for the performance evaluation of each individual we will show that the optimized parameters are robust with respect to noisy signals.

## 5 Recognition Performance

In order to evaluate the performance of the system, a database was built using 25 very frequent monosyllabic words extracted from the DARPA Resource



**Fig. 4.** Improvement of the recognition performance using an evolutionary algorithm to tune the parameters, compared to manual tuning one layer after the other. The spectrograms are scaled using a linear interpolation.

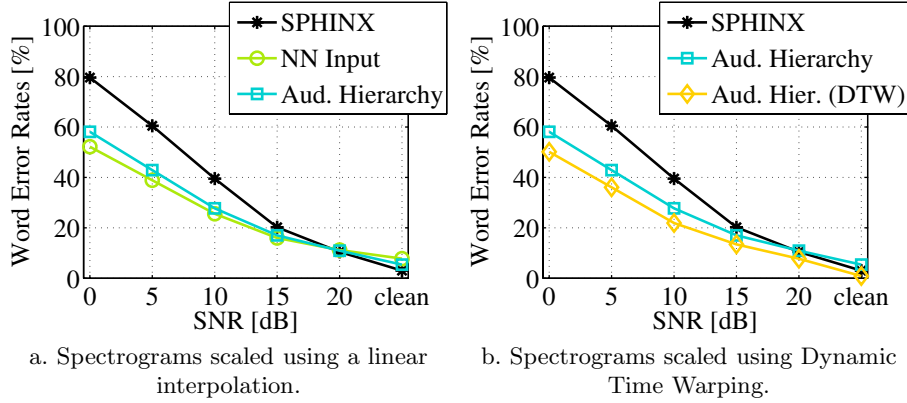
Management (RM) database. Isolated monosyllabic words have been chosen in lack of a syllable segmented database with sufficient size. The words were segmented using forced-alignment. For each of the monosyllabic words we selected 140 occurrences from 12 different speakers (6 males and 6 females) from the speaker dependent part of the database. For training 70 repetitions of each word were used, 20 for the early stopping validation of the Syllable-Tuned Units and 50 for testing.

The performance of our system has been compared to the Sphinx-4 speech recognition system, an open source speech recognition system that performs well on the whole RM corpus [15]. MFCC features were used as front-end for the HMMs. 13 cepstral coefficients plus delta and double delta were computed using the default parameters of Sphinx. Cepstral Mean Normalization [16] has been used in order to improve the robustness of the MFCC features. SphinxTrain was employed to train triphones HMMs. Each model had 3 states without skip over states and each state used a mixture of 8 Gaussians. The Hidden Markov Models were trained on the segmented monosyllabic words.

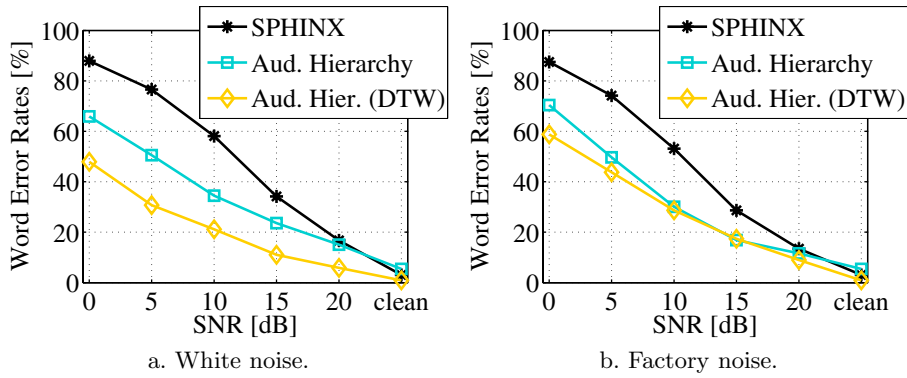
The robustness towards noise has been investigated by adding babble noise, white noise, and factory noise from the NOISEX database to the test database at different signal to noise ratios (SNR) while training was still performed on clean data.

Figure 4 illustrates the gain in performance on babble noise obtained using the evolutionary algorithm, compared to a manual tuning of the parameters one layer after the other. Following the notation introduced in [8], the optimal parameters given by the evolution strategy are  $\gamma_1 = 0.82$ ,  $\theta_1 = 2.66$ ,  $\sigma_1 = 3.16$  for the first layer and  $\gamma_2 = 0.84$ ,  $\theta_2 = 2.78$ ,  $\sigma_2 = 1.87$  for the second layer, when linear interpolation is used to scale the signals. Using a DTW, the optimal set of parameters is  $\gamma_1 = 0.99$ ,  $\theta_1 = 0.32$ ,  $\sigma_1 = 4$  for the first layer and  $\gamma_2 = 0.89$ ,  $\theta_2 = 0.99$ ,  $\sigma_2 = 1.93$ . As can be seen, the performance increased due to the optimization at all SNR levels. With clean speech we observe an improvement





**Fig. 5.** Comparison of the Word Error Rates (WER) between the proposed system and Sphinx-4 in the presence of babble noise



**Fig. 6.** Comparison of the Word Error Rates (WER) between the proposed system and Sphinx-4 in the presence of white and factory noise

from 6.72% to 5.44% (19% relative). The largest improvement was achieved at 15 dB SNR from 30.72% to 17.04% (44.5% relative).

Figure 5 summarizes the performance of both Sphinx-4 and the proposed system in presence of a babble noise. To measure the baseline similarities of the image ensemble, we also give the performance of a nearest neighbor classifier (NN) that matches the test data against all available training "views". An exhaustive storage of examples is, however, not a viable model for auditory classification. With clean signals, the STUs show better generalization capabilities and perform better than a nearest neighbor on the input layer (Fig. 5 a.). For noisy signals, the STUs are slightly worse, however, at a strong reduction of representational complexity.

With a simple linear time scaling our system only outperforms Sphinx-4 in noisy conditions but shows inferior performance on clean data. When Dynamic Time Warping is used to properly scale the signals, the STUs improve the already good performance obtained directly after the preprocessing in all the cases and our system outperforms Sphinx-4 even for clean signals (Fig. 5 b.). With clean data Sphinx obtains a 3.1% Word Error Rate (WER), our system achieves 0.9% WER with the DTW and 5.4% without the DTW. Figure 6 shows that the performance is very similar when adding white or factory noise.

## 6 Discussion and Summary

In this paper, we presented a novel approach to speech recognition interpreting spectrograms as images and deploying a hierarchical object recognition system. To optimize the main free parameters of the system, we used an evolutionary algorithm which allows us to quickly change the system without the need for manual parameter tuning.

We could show that our system performs better than a state of the art system in noisy conditions even when we applied a simplistic linear scaling of the input for time alignment. When we aligned the current utterance with the DTW to a known representation in an optimal non-linear way, we obtained better than state of the art results for all cases tested. However, in its current form the DTW makes use of information not available in real situations.

From this we conclude that our architecture and the underlying features are more robust against noise than the commonly used mel frequency cepstral coefficients (MFCCs). This robustness against noise is very important for real world scenarios which are usually characterized by significant background noise and variations in the recording conditions. A similar robustness was also observed for visual recognition in clutter scenes [8].

Our comparison between the linear scaling and the DTW shows that the performance of the model could be significantly improved by better temporal alignment. We therefore consider methods for improving this alignment as interesting future research directions.

The complexity of our recognition task is very low. Therefore, it remains an open question how our system will scale to more complex tasks. We can expect that our system generalizes well to larger vocabulary. In fact, the high dimensionality and the sparseness of the vector space at the output of the combination layer should allow to train STUs for a large number of syllables.

In order to process continuous speech, syllable segmentation is required. One way to obtain this segmentation is to implement a syllable segmentation system prior to the recognition. This would allow to keep the advantages of the recognition hierarchy: its fast implementation and the capacity to train or update STUs on the fly. Another possibility is to use the architecture as a front-end for Hidden Markov Models similarly to [17].

## References

1. Lippmann, R.: Speech recognition by machines and humans. *Speech Communication* 22(1), 1–15 (1997)
2. Sur, M., Garraghty, P., Roe, A.: Experimentally induced visual projections into auditory thalamus and cortex. *Science* 242(4884), 1437–1441 (1988)
3. Shamma, S.: On the role of space and time in auditory processing. *Trends in Cognitive Sciences* 5(8), 340–348 (2001)
4. Chih, T., Ru, P., Shamma, S.: Multiresolution spectrotemporal analysis of complex sounds. *Journal of the Acoustical Society of America* 118, 887–906 (2005)
5. Elhilali, M., Shamma, S.: A biologically-inspired approach to the cocktail party problem. In: *Proc. ICASSP*, vol. 5, pp. 637–640 (2006)
6. Mesgarani, N., Slaney, M., Shamma, S.: Discrimination of speech from non-speech based on multiscale spectro-temporal modulations. *IEEE Transactions on Speech and Audio Processing*, 920–930 (2006)
7. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. *Nature Neuroscience* 2, 1019–1025 (1999)
8. Wersing, H., Körner, E.: Learning optimized features for hierarchical models of invariant recognition. *Neural Computation* 15(7), 1559–1588 (2003)
9. Kleinschmidt, M., Gelbart, D.: Improving word accuracy with gabor feature extraction. In: *ICSLP*, Denver (2002)
10. Slaney, M.: An efficient implementation of the Patterson-Holdsworth auditory filterbank. Technical report, Apple Computer Co, Technical report #35 (1993)
11. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research* 5, 1457–1469 (2004)
12. Schneider, G., Wersing, H., Sendhoff, B., Körner, E.: Evolutionary optimization of a hierarchical object recognition model. *IEEE Transaction on Systems, Man and Cybernetics. Part B: Cybernetics* 35(3), 426–437 (2005)
13. Schwefel, H.P.: *Evolution and Optimum Seeking*. John Wiley and sons, New York (1995)
14. Bäck, T.: *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, Oxford (1996)
15. Walker, W., Lamere, P., Kwok, P.: Sphinx-4: A flexible open source framework for speech recognition. Technical report, Sun Microsystems Inc. (2004)
16. Liu, F.H., Stern, R.M., Huang, X., Acero, A.: Efficient cepstral normalization for robust speech recognition. In: *HLT 1993: Proceedings of the workshop on Human Language Technology*, Morristown, NJ, USA, Association for Computational Linguistics, pp. 69–74 (1993)
17. Meyer, B., Kleinschmidt, M.: Robust speech recognition based on localized, spectro-temporal features. In: *Elektronische Sprachsignalverarbeitung (ESSV)* (2003)