A Hierarchical Model for Syllable Recognition

Xavier Domont[1,2], Martin Heckmann[1], Heiko Wersing[1],
Frank Joublin[1], Christian Goerick[1]

[1]Honda Research Institute Europe, 63073 Offenbach/Main, Germany
{martin.heckmann, heiko.wersing}@honda-ri.de

[2]Technische Universität Darmstadt, 64283 Darmstadt, Germany
xavier.domont@rtr.tu-darmstadt.de

Abstract.
Inspired by recent findings on the similarities between the primary auditory and visual cortex we propose a neural network for speech recognition based on a hierarchical feedforward architecture for visual object recognition. When using a Gammatone filterbank for the spectral analysis the resulting spectrograms of syllables can be interpreted as images. After a preprocessing enhancing the formants in the speech signal and a length normalization, the images can than be fed into the visual hierarchy. We demonstrate the validity of our approach on the recognition of 25 different monosyllabic words and compare the results to the Sphinx-4 speech recognition system. Our hierarchical model achieves an improvement for high noise levels.

1 Introduction

In recent years significant similarities between the primary auditory and visual cortex have been revealed. Sur demonstrated in 1988 that, when their auditory cortex was fed with visual input, newborn ferrets developed some visual capacities in the auditory cortex [1]. More recently Shamma unveiled that the time-frequency receptive fields in the primary auditory cortex of ferrets show strong similarities to those of the visual cortex [2]. They are selective to modulations in the time-frequency domain and have Gabor-like shapes. These receptive fields have been modeled by Chin [3] and used for source separation [4] and speech detection [5]. Gabor-like filters have been used extensively in object recognition systems [6, 7].

The above findings motivated us to develop a system for speech recognition in strong resemblance to a hierarchical object recognition architecture. It is based on a feedforward neural network initially developed by Wersing and Körner for object recognition [7]. Our aim is to overcome the limitations of conventional speech recognition systems which substantially lack robustness. In our system we use syllables as speech units. Syllables being the basic units for speech production and showing less co-articulatory effects across their boundaries, we believe that they are the adequate speech units for a biologically-inspired system. Moreover the syllable segmentation required for the training of the system seems biologically plausible for speech acquisition.

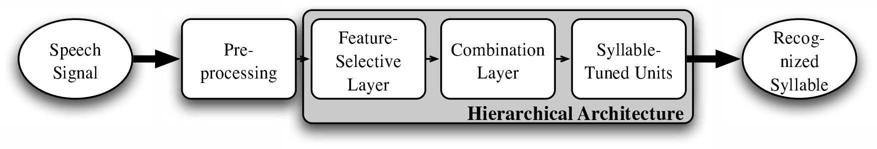

Fig. 1: Overview of the system.

In the following sections the building blocks of the system are detailed (Fig. 1). We then 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 aims at retaining only the phonetic information from the speech signal and removing speaker and recording specific parts. The formants, the resonances of the vocal tract, convey the main phonetic information. Therefore their trajectories are enhanced in the spectrogram. The resulting spectrogram can then be fed as "image" in the hierarchical recognition architecture.

The first step in the preprocessing is the application of a Gammatone filterbank which models the response of the basilar membrane in the human inner ear. The signal's sampling frequency is 16 kHz. The filterbank has 128 channels ranging from 80 Hz to 8 kHz. The spectrogram of the signal is calculated via rectification and low-pass filtering of the Gammatone filterbank response (Fig. 2 left). To compensate the influence of the speech excitation signal, the high frequencies are emphasized by +6 dB per octave resulting in a flattened spectrogram (Fig. 2 center). Next, the formant frequencies are enhanced by filtering along the channel axis using mexican-hat filters (Fig. 2 right). 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 makes the approximation 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 this recognition task. In order to also assess the performance of the recognition hierarchy independent of this non-linear scaling we also applied the Dynamic Time Warping (DTW) method on the spectrograms. For the DTW we selected one single repetition of a syllable and warped all the other repetitions to it. 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

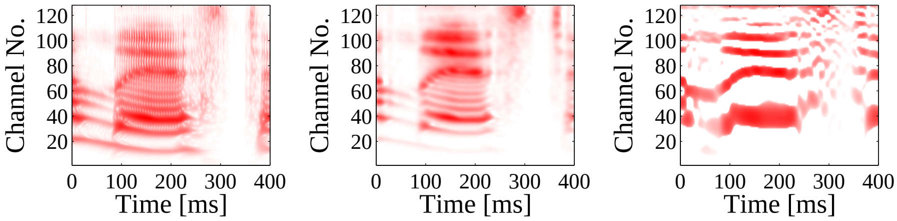

Fig. 2: Overview of the preprocessing step for the word "list" spoken by an american woman. Left: response of the basilar membrane. Center: after a low-pass filtering over the time and a preemphasis. Right the harmonic structure has been removed using a filtering along the frequency axis.

128×128, i.e. 128 time frames over 128 frequency channels. Note, however, that the application of DTW requires an hypothesis on the syllable to be known. Thus cannot easily be applied in a real recognition setting. Further work will focus at releasing this constraint.

3 The recognition hierarchy

The preprocessed two-dimensional spectrogram is from now on considered as an image and feeds a feedforward architecture initially aimed at object recognition. However, the structure of spectrograms differs from the structure of images taken from objects and, keeping the overall layout of the network described in [7], 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) mechanism and a pooling. The preprocessed spectrogram is firstly filtered by eight different Gabor-like filters. The purpose of these filters is to extract local features from the spectrogram. In [7] 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 in the system. The pooling performs a downsampling of the spectrogram by four in both time and frequency directions. The feature-selective layer transforms the 128×128 original spectrogram to eight 32×32 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 mechanism and a pooling. These combination

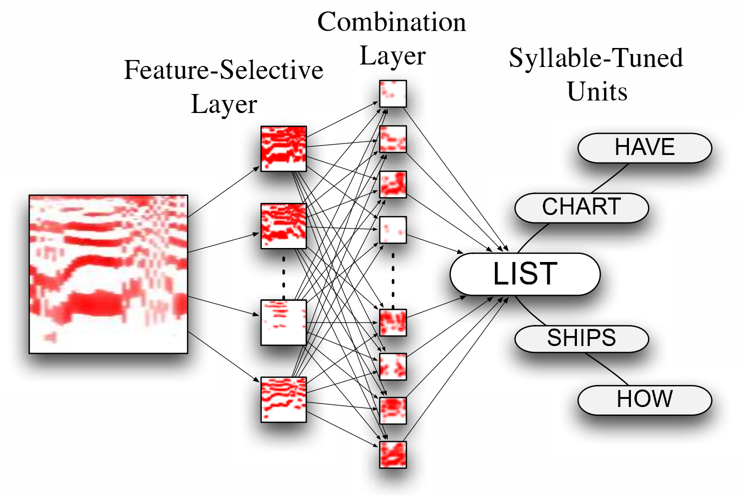

Fig. 3: The system is based on a feedforward architecture with weight-sharing and a succession of feature sensitive matching and pooling stages. It comprises three stages arranged in a processing hierarchy.

cells are learned using the non-negative sparse coding method (NNSC) as in [7], 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 β, 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. The NNSC is therefore expected to learn complex and global features appearing in the data. An exhaustive description of this method can be found in [8].

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

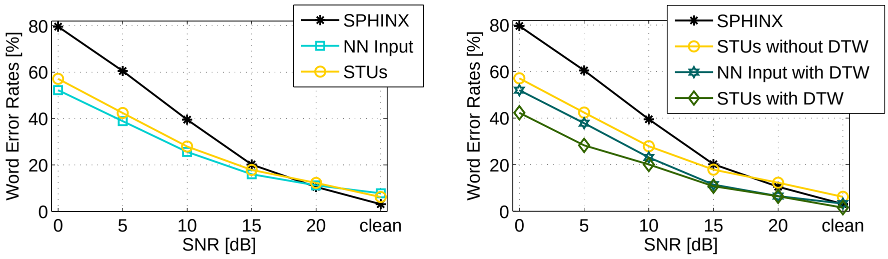

Fig. 4: Comparison of the Word Error Rates (WER) between the proposed system and Sphinx-4 in the presence of babble noise. Left: Recognition performance at the different layers of the hierarchy. Right: Recognition performance when Dynamic Time Warping is used to scale the signals.

reference to the View-Tuned Units used in [6] and [7].

4 Recognition performances

In order to evaluate the performance of the system, a database was built using 25 very frequent monosyllabic words extracted from the DARPA Resource 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 database part. 70 repetitions of each word were used for training, 20 for the early stopping validation of the Syllable-Tuned Units and 50 for testing.

The parameters of the WTM competitions and poolings were optimized using a raster method. Following the notations introduced in [7], $\gamma_1 = 0.85$, $\theta_1 = 3$, $\sigma_1 = 2.5$ for the first layer and $\gamma_2 = 0.8$, $\theta_2 = 1.1$, $\sigma_2 = 2$ for the second layer.

The performance of our system has been compared to the Sphinx-4 speech recognition system, an open source speech recognition system well performing on the whole RM corpus [9]. The Hidden Markov Models for Sphinx were trained only on the segmented monosyllabic words. The robustness towards noise has been investigated adding babble noise to the test database at different signal to noise ratios (SNR) while training was still performed on clean data. Fig. 4 summarizes the performance of both Sphinx-4 and the proposed system.

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. 4 left). 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 for low SNRs, but, when Dynamic Time Warping is used to proper scale the signals, the STUs improve the already good performances obtained directly after the preprocessing in all the cases and outperforms Sphinx-4 even for clean signals

(Fig. 4 right). With clean data Sphinx obtains a 3.1% Word Error Rate (WER), our system with the DTW achieves 1.5% WER with the DTW and 6.2% without the DTW.

5 Discussion

In this paper we presented a novel approach to speech recognition interpreting spectrograms as images and deploying a hierarchical object recognition system. We could show that such a 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.

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 in 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 [7].

Our comparison to using 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.

References

[1] M Sur, PE Garraghty, and AW Roe. Experimentally induced visual projections into auditory thalamus and cortex. *Science*, 242(4884):1437–1441, 1988.

[2] S. Shamma. On the role of space and time in auditory processing. *Theoretical Comput. Sci.*, 5(8):340–348, 2001.

[3] T. Chih, P. Ru, and S. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *Journal of the Acoustical Society of America*, 118:887–906, 2005.

[4] M. Elhilali and S. Shamma. A bilogically-inspired approach to the cocktail party problem. In *Proc. ICASSP*, volume 5, pages V–637–640, 2006.

[5] N. Mesgarani, M. Slaney, and S. Shamma. Discrimination of speech from non-speech based on multiscale spectro-temporal modulations. *IEEE Transactions on Speech and Audio Processing*, pages 920–930, 2006.

[6] M. Riesenhuber and T. Poggio. Hierachical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.

[7] H. Wersing and E. Körner. Learning optimized features for hierarchical models of invariant recognition. *Neural Computation*, 15(7):1559–1588, 2003.

[8] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[9] W. Walker, P. Lamere, and P. Kwok. Sphinx-4: A flexible open source framework for speech recognition. Technical report, Sun Microsystems Inc., 2004.