Combining Auditory Inspirations and Hierarchical Feature Extraction for Robust Speech Recognition

Martin Heckmann[1], Xavier Domont[1,2], Frank Joublin[1], Christian Goerick[1]
[1] *Honda Research Institute Europe, 63073 Offenbach/Main, Germany, Email: firstname.lastname@honda-ri.de*
[2] *Techn. Universität Darmstadt, Regelungsth. u. Robotik, Germany, Email: xavier.domont@rtr.tu-darmstadt.de*

Abstract

We present speech features inspired by the processing in the auditory periphery and the receptive fields found in the auditory cortex. They have a hierarchical organization and jointly evaluate variations in the spectro-temporal domain. This is why we termed them Hierarchical Spectro-Temporal (HIST) features.

For their calculation we apply a Gammatone filterbank to transform the signal into the spectral domain. In a preprocessing based on local competition mechanisms we enhance the formants in the spectrogram. A set of filters learned via ICA (Independent Component Analysis) captures local variations in the spectrogram and constitutes the first layer of the hierarchy.

In the second layer these local variations are combined to form larger receptive fields learned via Non Negative Sparse Coding. The dimensionality of the resulting features is reduced via the application of a Principal Component Analysis (PCA) and then fed into a Hidden Markov Model (HMM).

We evaluated the performance of these features in a continuous digit recognition task in a variety of different noise conditions, similar to the Aurora task. Our results show, especially in combination with RASTA features, a significant performance improvement in noise.

Introduction

Already for a long time the process of human speech perception serves as a role model in the development of machine recognition (e. g. RASTA-PLP [1]). Here, we present features which take their inspiration not from psychoacoustic but neurophysiological data. Shamma showed that the primary auditory cortex of young ferrets has a spectro-temporal organization, i. e. the receptive fields are selective to modulations in the time-frequency domain and, as in the visual cortex, have Gabor-like shapes [2]. However, traditionally speech features rely only on spectral representations. Such spectro-temporal features were already used for speech recognition [3, 4, 5], speech detection [6, 7], and source separation [8].

Justified by the found analogies between the visual and auditory cortex in mammals, we developed speech features in strong similarity to the visual object recognition system described in [9]. Its main features are the hierarchical organization in three layers and the unsupervised learning of the receptive fields on the first and second layer. We termed the speech features we derived thereof as Hierarchical Spectro-Temporal (HIST) features and used them a front-end to Hidden Markov Models (HMMs) [10]. In this paper we report further improvements of these features and tests on a continuous digit recognition task.

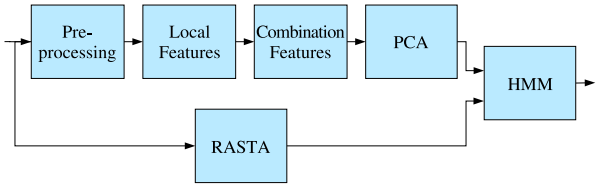

Figure 1: Overview of the feature extraction process.

In the following section, the computation and enhancement of the spectrograms are described. The calculation of the HIST features from the spectrograms is explained in the section after that (see Fig. 1 for an overview of the process). Finally, the performance of the HIST features is evaluated, especially in respect to RASTA-PLP features, in the before last section.

Preprocessing

The spectrograms of the speech signals were computed using a Gammatone filter-bank. We used an Infinite Impulse Response (IIR) implementation of the Gammatone filter-bank [11] having 128 channels ranging from 80 Hz to 8 kHz at a sampling rate of 16 kHz. The spectrograms are obtained by rectification and low-pass filtering of the filter-bank response. The sampling rate of the spectrograms was then reduced to 400 Hz.

Formant enhancement

The remaining preprocessing steps enhance the formants in the spectrogram. Via a preemphasis of +6 dB/oct. the influence of the speech excitation signal was compensated for. Next, we used a set of Mexican Hat filters along the frequency axis to remove the harmonic structure of the spectrograms and form peaks at the formant locations. The size of the filter kernels was chosen constant on a linear frequency axis. Due to the logarithmic arrangement of the center frequencies in the Gammatone filterbank in the implementation the size of the kernels varied accordingly. Additionally, the shapes of the filters were adapted to the nonlinear frequency spacing, i. e. the lower part of the filter is wider than the higher part. A second Mexican Hat filter with smaller kernel sizes for lower frequencies thinned the resulting formant tracks. Figure 2 shows the original spectrogram and the result of the formant enhancement of the digit "one" spoken by a male

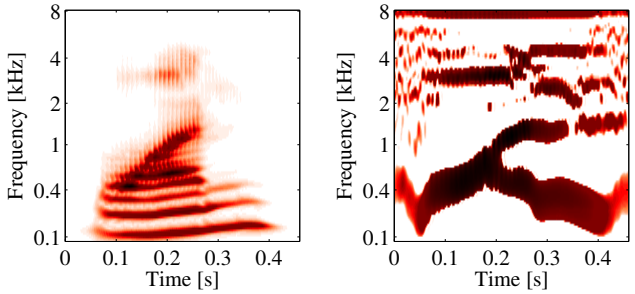

Figure 2: Original (a) and enhanced spectrogram (b) of the digit "one" spoken by a male speaker

American speaker in clean conditions.

Hierarchical spectro-temporal features

The feature extraction is build upon two hierarchical levels: The first level extracts local features and the second level integrates them to more complex features, spanning the whole frequency range. We learned these features solely on the test set of the database.

First stage: Extraction of local features

The extraction of local features on the first level was performed via a 2D filtering with a set of n_1 receptive fields $\mathbf{w}_1^l$, taking the absolute value of the response:

$$q_1^l(t,f) = \left|\left(\mathbf{S} * \mathbf{w}_1^l\right)(t,f)\right|, \tag{1}$$

where the responses $\mathbf{q}_1^l$ of each neuron had the same size as the input spectrogram $\mathbf{S}$. Here and in the following we interpret the spectrogram as a 2D image.

These $n_1 = 8$ receptive fields have been learned using Independent Component Analysis (ICA) on 3500 randomly selected local 16×16 patches of the enhanced spectrograms taken from the training set.

For a given point (t, f) in the spectrogram, the activity $q_1^l(t,f)$ of the lth neuron reveals how close a local patch of $\mathbf{S}$ centered in (t, f) is to the pattern l. For each local patch only the highest correlated patterns are of interest. Therefore, we performed a Winner-Take-Most (WTM) competition which inhibited the response of the less active neurons at the position (t, f):

$$r_1^l(t,f) = \begin{cases} 0 & \text{if } \frac{q_1^l(t,f)}{M(t,f)} < \gamma_1 \text{ or } M(t,f) = 0 \\ \frac{q_1^l(t,f) - \gamma_1 M(t,f)}{1-\gamma_1} & \text{else,} \end{cases} \tag{2}$$

where $M(t,f) = \max_k q_1^k(t,f)$ is the maximal value at position (t, f) over the eight neurons and $0 \le \gamma_1 \le 1$ is a parameter controlling the strength of the competition [9].

Furthermore, a nonlinear transformation including a threshold θ_1 was applied on all the $r_1^l(t,f)$:

$$s_1^l(t,f) = H(r_1^l(t,f) - \theta_1), \tag{3}$$

where $H(x)$ is the Heaviside step function.

After smoothing with a 2D Gaussian filter $\mathbf{g}_1$ the resolution of the images $\mathbf{s}_l(t,f)$ was reduced by a factor of four in both frequency and time dimension

$$c_1^l(t,f) = \left(\mathbf{s}_1^l * \mathbf{g}_1\right)(4t, 4f) \tag{4}$$

yielding 32 frequency channels and a sampling rate of 100 Hz.

Second stage: Extraction of combination features

Each of the n_2 combination patterns is composed of n_1 receptive fields $\mathbf{w}_{2,l}^k$, i. e. one for each of the neurons in the previous stage. The coefficients of these receptive fields are non negative and span all frequency channels. Similarly to (1) the activity $q_2^k(t)$ of the kth neuron at the time t is given by:

$$q_2^k(t) = \sum_{l=1}^{n_1} \left(\mathbf{c}_1^l * \mathbf{w}_{2,l}^k\right)(t,f). \tag{5}$$

As the combination patterns span the whole frequency range the response of the neurons does not depend on f anymore. This means that, by computing the convolution, the patterns $\mathbf{w}_{2,l}^k$ are only shifted in the time direction. Note that the absolute value is not required in (5) as both the $\mathbf{c}_1^l$ and the $\mathbf{w}_{2,l}^k$ are non-negative.

The combination patterns were also learned in an unsupervised manner using Non-Negative Sparse Coding (NNSC) [12]. NNSC differs from Non-Negative Matrix Factorization (NMF) by the presence, in the cost function (6), of a sparsity enforcing term 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.

We cut out patches of length $\Delta = 20$ ms of the first layer activations $\mathbf{c}_1^l$. From these patches we learned $n_2 = 50$ combination features by minimizing the following cost function [9]:

$$E = \sum_p \|\mathbf{P}^p - \sum_{k=1}^{n_2} \alpha_k^p \mathbf{w}_2^k\|^2 + \beta \sum_p \sum_{k=1}^{n_2} |\alpha_k^p|, \tag{6}$$

where $\mathbf{P}^p$ is a tensor representing the n_1 layers of the p-th patch, the $\mathbf{w}_2^k$ are n_2 non-negative tensors each of them containing the n_1 receptive fields $\mathbf{w}_{2,l}^k$, the α_k^p are nonnegative reconstruction factors, and β is a parameter allowing to control the sparsity of the learned features.

This yielded $n_2 = 50$ features $\left(q_2^1(t), \ldots, q_2^{n_2}(t)\right)^T$ at a feature rate of 100 Hz. Delta (resp. double-delta) features were computed using a 9th order FIR lowpass (resp. bandpass). The dimensionality of the feature vectors was then reduced from 150 to 39 using Principal Component Analysis (PCA) learned on the training set.

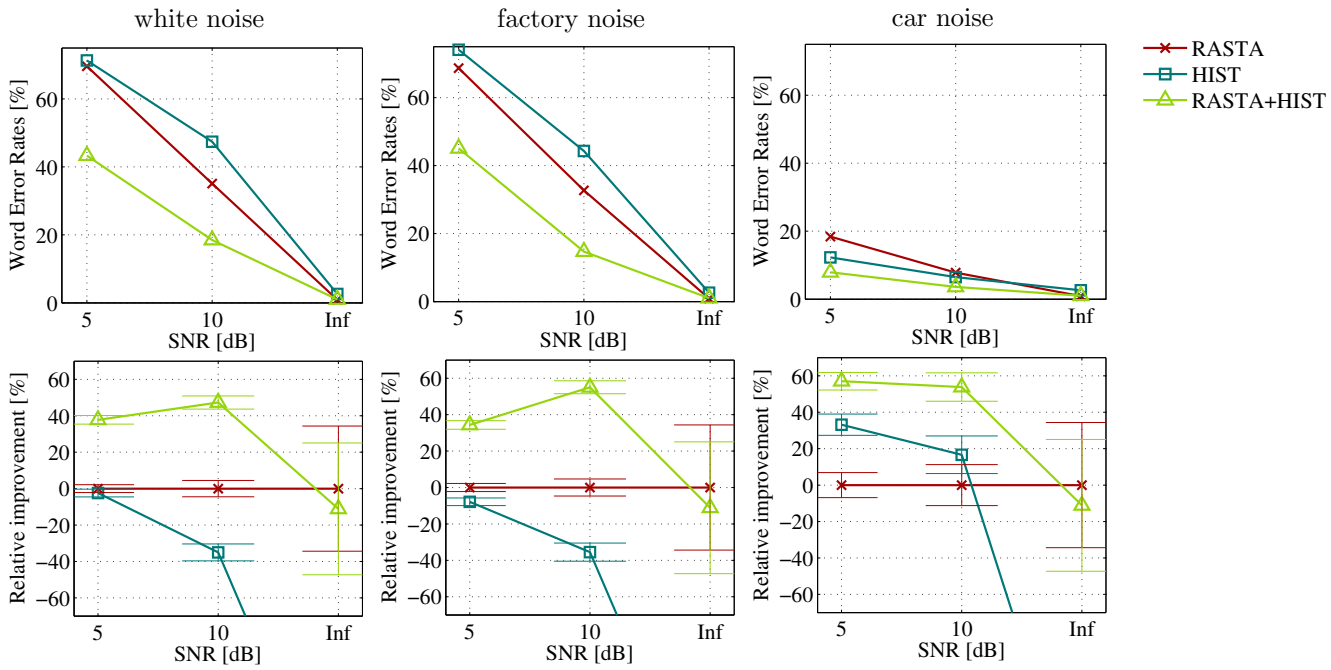

Figure 3: Recognition scores obtained for RASTA-PLP , HIST , and combined RASTA-PLP -HIST features on the complete TIDigits dataset. Noise type added and SNR level are indicated in the plots. The upper part shows the absolute errors, the lower part the errors relative to those obtained using only RASTA-PLP features. The bars in the lower part indicate the 95% confidence interval.

Recognition performance

The recognition task

The following recognition experiments were performed on the TIDigits corpus [13]. It contains continuously uttered digits from 326 speakers of different age and gender. The speakers in the evaluation set are different from those in the training set. We mixed the utterances of the test database with additive noise in a similar way as in the Aurora-2 framework [14] . The main differences were:

- We downsampled signals to 16 kHz instead of 8 kHz.
- When mixing the signals with noise using FaNT [15] we used the G.712 only for the noise and signal level estimation, i. e. the obtained signals have no channel distortions.
- A reduced set of noise types from the Noisex database [16] were used. White, Factory, and Car at *Signal to Noise Ratio (SNR)* levels of 10 and 5 dB.

The Hidden Markov Models were trained on clean signals with HTK [17] using the same parameters as in the Aurora-2 framework [14]. Whole word HMMs containing 16 states without skip transitions and a mixture of 3 Gaussians with a diagonal covariance matrix per state were used.

Comparison with State of the Art features

To assess the performance of the proposed features, we compared our results to RASTA-PLP features. We used an order of 14 for the linear prediction and also delta and double-delta coefficients. In all cases the HMMs were trained on clean signals. Additionally, we combined the HIST and RASTA-PLP features to study their complementarity.

	RASTA-PLP	HIST	RASTA-PLP+HIST
Absolute	0.9	2.6	1.0
Relative	0±34	−189±58	−11±36

Table 1: Comparison of results between RASTA-PLP , HIST , and HIST +RASTA-PLP . Values are given as percent absolute Word Error Rates and relative Word Error Rates (with 95% confidence interval), relative to RASTA-PLP features.

In Fig. 3 the recognition scores when adding white noise, factory noise, and car noise at different SNR levels are given. As can be seen the HIST features perform similar or worse than RASTA-PLP for most cases. However, when combined with RASTA-PLP features we see clear improvements in the range of $40 - 60\%$ relative. This is best seen in the lower part of Fig. 3 where the word error rates are given relative to the ones obtained by RASTA-PLP features. For clean speech performance of HIST combined with RASTA-PLP is a little worse than RASTA-PLP alone. Yet these results are due to the very small error rates not statistically significant. (the 95% confidence intervalls were calculated according to [18]). These small differences in error rate and the small reference value in the calculation of the relative error are also the reason why in the lower part of the plot the relative errors are outside of the plot. The corresponding values are given in Tab. 1.

Already by themselves the HIST features seem to be able to cope quite well with car noise. A reason for this could be that car noise has mainly low frequency components. These are partly filtered out by the

Gammatone filterbank already as it has a low cut-off frequency of 80 Hz. Additionally, in our preprocessing the different frequency channels are dealt with more or less independently. Hence disturbances at low frequencies do not affect higher frequencies. This is in contrast to RASTA-PLP features where disturbances can effect all poles during the calculation of the predictor coefficients.

This independence of the frequency channels could also contribute to the improvements we see when combining RASTA-PLP and HIST. However, we assume that rather the better representation of formant transitions via the spectro-temporal processing is the main cause for this improvement.

Discussion & Summary

In this paper we investigated our previously presented HIST features further [10]. We kept the organization into two hierarchical layers: the first detecting local spectro-temporal variations and the second combining them into features covering a larger section of the spectrogram. On both levels the features were learned in an unsupervised way, thereby allowing in principle a continuous learning.

For the evaluation we used the TIDigits database which contains continuously uttered digits from a wide range of speakers. The setup was similar to Aurora-2 with the main restriction that we performed the evaluation only in a subset of the noise types and SNR levels used there.

We previously observed that HIST features show good performance at low SNRs. Here we could not really reproduce this as we only investigated good to moderate SNR levels. Nevertheless, one can see that the difference between RASTA-PLP and HIST features is much smaller for 10 dB than for 5 dB. When combined with RASTA-PLP features a significant improvement could be obtained compared to RASTA-PLP alone. We previously observed this behavior when we performed tests on the isolated digit part of TIDigits with a much more comprehensive set of SNR levels. In light of these results we are very confident that the improvement obtained in the combintation of RASTA-PLP and HIST features scales up to more complex recognition tasks.

In summary, we could show that our HIST features deliver complementary information to conventional spectral features. In our opinion, it is the spectro-temporal information, i. e. the transitions, which is better modeled via the HIST features and responsible for the improvement.

References

[1] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans Speech and Audio Proc.*, vol. 2, no. 4, pp. 578–589, 1994.

[2] S. Shamma, "On the role of space and time in auditory processing," *Trends in Cognitive Sciences*, vol. 5, no. 8, pp. 340–348, 2001.

[3] M. Kleinschmidt, *Robust Speech Recognition Based on Spectro-temporal Processing*, Ph.D. thesis, Universität Oldenburg, 2002.

[4] B. Meyer and B. Kollmeier, "Optimization and evaluation of gabor feature sets for asr," in *Proc. Interspeech*, 2008.

[5] Nelson Morgan Sherry Y. Zhao, "Multi-stream spectro-temporal features for robust speech recognition," in *Proc. INTERSPEECH*, 2008.

[6] N. Mesgarani, M. Slaney, and SA Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 14, no. 3, pp. 920–930, 2006.

[7] Tomaso Poggio Tony Ezzat, "Discriminative word-spotting using ordered spectro-temporal patch features," in *Proc. SAPA*, 2008.

[8] M. Elhilali and S. Shamma, "A Biologically-Inspired Approach to the Cocktail Party Problem," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 2006.

[9] H. Wersing and E. Körner, "Learning Optimized Features for Hierarchical Models of Invariant Object Recognition," *Neural Computation*, vol. 15, no. 7, pp. 1559–1588, 2003.

[10] X. Domont, M. Heckmann, F. Joublin, and C. Goerick, "Hierarchical sectro-temporal features for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, Las Vegas, Nevada, 2008, pp. 4417–4420, IEEE.

[11] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Tech. Rep., Apple Computer Co., 1993, Technical report #35.

[12] P.O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[13] R. Leonard, "A Database for Speaker-independent Digit Recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, 1984, vol. 9.

[14] D. Pearce and H.G. Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," in *Int. Conf. on Spoken Lang. Proc.* 2000, ISCA.

[15] G. Hirsch, "FaNTFiltering and Noise Adding Tool," Tech. Rep., Niederrhein University of Applied Sciences, 2005.

[16] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[17] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University, 1995.

[18] M. Heckmann, *Adaptive Datenfusion für die Audio-visuelle Spracherkennung*, Shaker, 2003.