

A Closer Look on Hierarchical Spectro-Temporal Features (HIST)

**Martin Ernst Heckmann, Xavier Domont, Frank Joublin,
Christian Goerick**

2008

Preprint:

This is an accepted article published in Proc. INTERSPEECH 2008. The final authenticated version is available online at: [https://doi.org/\[DOI not available\]](https://doi.org/[DOI not available])

A Closer Look on Hierarchical Spectro-Temporal Features (HIST)

Martin Heckmann¹, Xavier Domont^{1,2}, Frank Joublin¹, Christian Goerick¹

¹Honda Research Institute Europe GmbH, D-63073 Offenbach am Main, Germany

{martin.heckmann, frank.joublin, christian.goerick}@honda-ri.de

²Technische Universität Darmstadt, D-64283 Darmstadt, Germany

xavier.domont@rtr.tu-darmstadt.de

Abstract

Speech recognition robust against interfering noise remains a difficult task. We previously presented a set of spectro-temporal speech features which we termed Hierarchical Spectro-Temporal (HIST) features showing improved robustness, especially when combined with RASTA-PLP. They are inspired by the receptive fields found in the mammalian auditory cortex and are organized in two hierarchical levels. A set of filters learned via ICA captures local variations 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.

In this paper we introduce a non-linear smoothing along the time axis of the spectrograms at the input to the hierarchy and, additionally, a more thorough performance analysis on an isolated and a continuous digit recognition task. The results show that the combination of HIST and RASTA-PLP features yields improved recognition scores in noise.

Index Terms: Spectro-temporal, auditory, robust speech recognition, non-linear smoothing

1. Introduction

Human speech perception performance, especially in noisy conditions, remains unmatched by machine recognition. Driven by this, different feature extraction methods based on psychoacoustic models were developed (e.g. RASTA-PLP [1]). The recognition architecture we present here tries to go one step further in the sense that it takes its inspiration from observations on the organization of the primary auditory cortex of mammals.

Despite the spectro-temporal nature of speech, traditional speech features rely only on spectral representations. 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]. Such spectro-temporal features were already used for speech recognition [3], speech detection [4], and source separation.

We previously proposed a feed-forward neural network for isolated monosyllabic word recognition deploying spectro-temporal features. Justified by the found analogies between the visual and auditory cortex in mammals, it was inspired by the visual object recognition system described in [5]. Its main features are the hierarchical organization in three layers and the unsupervised learning of the receptive fields on the first and second layer. Based on this system we developed a set of Hierarchical Spectro-Temporal (HIST) features which we used as a front-end to Hidden Markov Models (HMMs) [6]. In this paper we introduce a non-linear smoothing of the spectrograms

prior to the calculation of the features and a more detailed performance analysis.

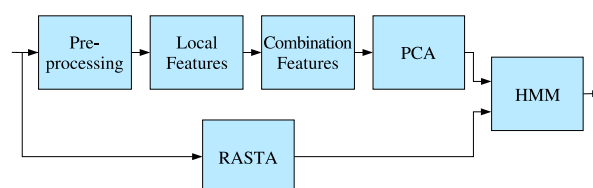


Figure 1: Overview of the feature extraction process.

In section 2, the computation and enhancement of the spectrograms, including the non-linear smoothing, are described. The calculation of the HIST features from the spectrograms is explained in section 3 (see Fig. 1 for an overview of the process). Finally, the performance of the HIST features is evaluated in respect to RASTA-PLP and MFCC features in section 4.

2. 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 [7] 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.

2.1. Non-linear smoothing

To suppress the remaining ripples in the spectrograms and, at the same time, maintain the onsets we introduced a non-linear smoothing. It acts in two modes. In the first mode the smooth envelope $x_s(k)$ rises with the signal envelope $x(k)$. When the signal envelope changes from the rising phase to a falling phase it changes its mode and now performs a smoothing of the envelope signal with a first order IIR filter. As soon as the envelope signal rises above the smooth signal the smoothing changes again in its rising phase:

$$x_s(k) = \begin{cases} 0, & k = 0 \\ x(k), & x_s(k-1) \leq x(k) \wedge k > 0 \\ (1 - 1/\tau) \cdot x_s(k-1) + 1/\tau \cdot x(k), & x_s(k-1) > x(k) \wedge k > 0. \end{cases} \quad (1)$$

As a consequence the onsets are conserved and the signal is only smoothed after the onsets.

2.2. 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 filter-bank 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

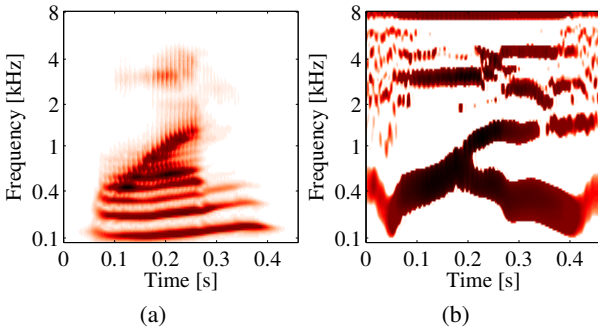


Figure 2: Original (a) and enhanced spectrogram (b) of the digit "one" spoken by a male speaker

and the result of the formant enhancement of the digit "one" spoken by a male American speaker in clean conditions.

3. 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.

3.1. 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:

$$q_1^l(t, f) = \left| \left(\mathbf{S} * \mathbf{w}_1^l \right) (t, f) \right|, \quad (2)$$

where the responses 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 l th 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) :

$$r_1^l(t, f) = \begin{cases} 0 & \text{if } \frac{q_1^l(t, f)}{M(t, f)} < \gamma_1 \text{ or } M(t, f) = 0 \\ \frac{q_1^l(t, f) - \gamma_1 M(t, f)}{1 - \gamma_1} & \text{else,} \end{cases} \quad (3)$$

where $M(t, f) = \max_k q_1^k(t, f)$ is the maximal value at position (t, f) over the eight neurons and $0 \leq \gamma_1 \leq 1$ is a parameter controlling the strength of the competition [5].

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

$$s_1^l(t, f) = H(r_1^l(t, f) - \theta_1), \quad (4)$$

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

After smoothing with a 2D Gaussian filter $\mathbf{g}_1$ the resolution of the images $s_1^l(t, f)$ was reduced by a factor of four in both frequency and time dimension

$$c_1^l(t, f) = \left(s_1^l * \mathbf{g}_1 \right) (4t, 4f) \quad (5)$$

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

3.2. 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 (2) the activity $q_2^k(t)$ of the k th neuron at the time t is given by:

$$q_2^k(t) = \sum_{l=1}^{n_1} \left(c_1^l * \mathbf{w}_{2,l}^k \right) (t, f). \quad (6)$$

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 (6) as both the 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) [8]. NNSC differs from Non-Negative Matrix Factorization (NMF) by the presence, in the cost function (7), 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 c_1^l . From these patches we learned $n_2 = 50$ combination features by minimizing the following cost function [5]:

$$E = \sum_p \|\mathbf{P}^p - \sum_{k=1}^{n_2} \alpha_k^p \mathbf{w}_2^k\|^2 + \beta \sum_p \sum_{k=1}^{n_2} |\alpha_k^p|, \quad (7)$$

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 $(q_2^1(t), \dots, q_2^{n_2}(t))^T$ 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.

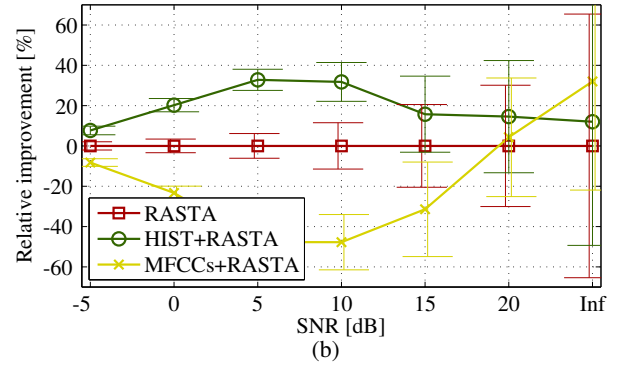
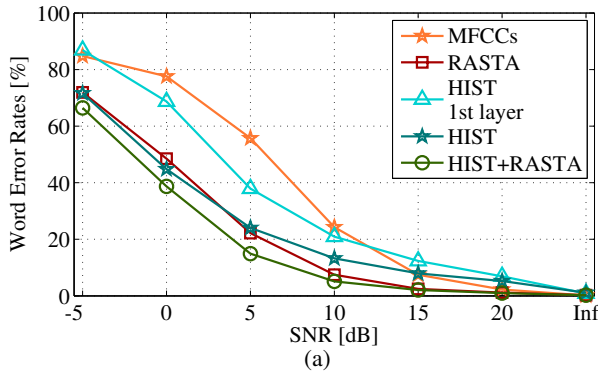


Figure 3: Comparison of the absolute (a) and relative performance on isolated digits with RASTA-PLP features as a baseline in (b). As noise factory noise was added. The bars in (b) mark the 95% confidence interval for the combination of HIST and RASTA-PLP features, MFCC and RASTA-PLP features, and the RASTA-PLP features alone.

4. Recognition performance

4.1. The recognition task

The time intensive processing of the current Matlab implementation restricted the main part of the evaluation to the isolated digit part of the TIDigits corpus [9]. However, some tests were also performed on the complete continuous digits part. If not indicated otherwise the following results were obtained on the isolated digits set. We mixed the utterances of the test database with additive noise in a similar way as in the Aurora-2 framework[10]. The differences were:

- We downsampled signals to 16 kHz instead of 8 kHz.
- When mixing the signals with noise using FaNT [11] we used the G.712 only for the noise and signal level estimation, i. e. the obtained signals have no channel distortions.
- Three types of noise from the Noisex database [12] were used: Babble, Factory, and Car.

Both, training and test set, contained 326 utterances for each of the 11 digits spoken by different speakers (boys, girls, women, and men). The speakers in the test set differed from those in the training set.

The Hidden Markov Models were trained on clean signals with HTK using the same parameters as in the Aurora-2 framework [10]. Whole word HMMs containing 16 states without skip transitions and a mixture of 3 Gaussians with a diagonal covariance matrix per state were used. For the isolated digits no model for pauses between words was used.

4.2. Comparison with State of the Art features

To assess the performance of the proposed features, we compared our results to MFCC and RASTA-PLP features. We used 12 MFCCs, without the zeroth coefficient but with the logarithmic frame energy plus the corresponding delta and double-delta coefficients. Cepstral Mean Subtraction was applied on the MFCCs. For the 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 different types of features to study their complementarity.

As can be seen in Fig. 3, in the presence of factory noise the performance of MFCCs decreases rapidly when the SNR falls below 15 dB. The RASTA-PLP features show a better performance and robustness than MFCCs, except for clean. The performance of the proposed HIST features decreases more

Smoothing	Noise Type				
	factory	babble	white	car	Ø
linear	-1.7±6.3	7.5±6.3	23.1±2.1	42.2±3.2	17.8
nonlinear	20.4±5.9	-1.3±6.8	28.2±2.0	47.2±3.0	23.7

Table 1: Average relative improvement (with 95% confidence interval) of the combination of RASTA-PLP and HIST features relative to RASTA-PLP features alone for the original linear smoothing and the proposed non-linear smoothing. Values were averaged over SNRs of $-5 \dots 20$ dB. Last column gives average over all noise types. The tests were performed on isolated digits.

smoothly when the noise level increases but they do not perform well on high SNRs and therefore only catch up on the RASTA-PLP features when the SNR is below 5 dB.

To investigate the effect of the hierarchical processing on the recognition we also added the recognition scores obtained when using only the first layer of the HIST architecture. For the calculation of these features we performed a DCT, kept only the first 50 values, and added delta and delta-delta features. As can be seen they perform not very well and the second level in the hierarchy clearly improves the results.

The best word error rates (except for clean) were obtained by concatenating the HIST features with RASTA-PLP features. In clean conditions the MFCCs performed best (0.17% WER, 6 errors). The differences between RASTA-PLP (0.25%, 8 errors) and the combination of RASTA-PLP and HIST features (0.22%, 7 errors) are very small and not statistically significant (calculated according to [13] with a confidence level of 95%). This can be better seen from Fig. 3(b) where we plotted the relative improvement of the HIST features combined with RASTA-PLP features against the RASTA-PLP features alone (including the limits of the 95% confidence interval).

In an additional test we investigated if the improvements we observed via the combination of HIST and RASTA-PLP features could also be obtained via a concatenation of MFCC and RASTA-PLP features. As can be seen from Fig. 3(b) this is not the case (at clean the performance of the combination is identical to MFCC features alone).

Tab. 1 gives a summary of the result for the combination of RASTA-PLP and HIST features, with a linear and a non-linear smoothing of the envelope. The other features (MFCC, MFCC + RASTA-PLP, and HIST alone) are omitted as they yielded negative values. As can be seen, except for babble noise, the average improvement with the non-linear smoothing is for all types of noise larger than 20%. The linear smoothing performs

	clean	white 10dB	white 5dB
RASTA-PLP	0.9±0.2	35.1±0.8	69.6±0.8
HIST linear+RASTA-PLP	1.0±0.2	17.4±0.7	39.0±0.9

Table 2: Comparison of results between RASTA-PLP and HIST with linear preprocessing combined with RASTA-PLP on complete TIDigits. Values are given as percent absolute Word Error Rates (with 95% confidence interval).

worse, especially in factory noise, but still yields considerable improvements for some cases. Overall the non-linear smoothing improved the results.

The use of an MVA post-processing [14] in an additional test significantly improved the performance of MFCC features at SNRs below 10 dB and to a lesser extent that of RASTA-PLP features. However, it did not influence the ranking in the performance of the features, especially the combination of HIST and RASTA-PLP features.

The tests reported so far were on a very simple task. To put our results on a more solid basis we performed some additional tests on the complete TIDigits database. It comprises 12549 utterances in the train and 12547 utterances in the test set of continuously spoken digits. We added white noise as described for the isolated digits. Due to the significant higher computational load we were only able to perform a very limited number of tests and only with the linear smoothing. All stages of the hierarchy were trained from scratch on the new dataset. The results are given in Tab. 2. As can be seen the general trend is identical to the one reported on the isolated digit task.

5. Discussion & Summary

In this paper we reported further improvements of our previously presented HIST features and performed some additional analysis. 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. The improvements we presented here were focused on the preprocessing of the spectrograms, namely the introduction of a non-linear smoothing of the spectrograms and some smaller changes.

We evaluated our features on a speaker-independent isolated digit recognition task. A similar setup as in Aurora-2 was used with 4 different types of additive noise, making the results, despite the small data set, quite conclusive. This is supported by the confidence intervals calculated for each measurement. As a continuous speech recognition task is certainly better suited to assess the performance of the features we also performed an additional test with limited noise types on a continuous digit recognition task. We abandoned the combination of HIST features of different window lengths as in [6] and used a single time window similar to that of RASTA-PLP and MFCC features instead. The very long time windows previously used harbored the risk of a learning of the complete digit.

Generally the HIST features showed good performance at low SNRs. When combined with RASTA-PLP features a significant improvement could be obtained compared to RASTA-PLP alone. MFCC features did only perform well in clean scenarios. A comparison with a combination of MFCC and RASTA-PLP features showed that, in contrast to MFCCs, the HIST features supply additional information. However, at high SNR levels HIST features were inferior to MFCC and RASTA-PLP features.

We attribute this to the significant downsampling we performed after the first level of our hierarchy and the insufficient preprocessing. The lack of assumptions on possible locations of the formants or on their number in our preprocessing might also be the cause that we did not see any improvements for the combination of HIST and RASTA-PLP features on babble noise.

Babble noise was also the only case where the nonlinear, onset conserving smoothing did not improve the performance of the HIST features compared to the linear smoothing. In the other cases we observed an improvement of $\approx 5\%$. Thereof we draw the conclusion that the better modeling of the onsets is beneficial.

The results obtained on the continuous digit task support our view that the improvement obtained via 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 and that this effect can not be replicated by simply concatenating MFCC and RASTA-PLP 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.

6. 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] N. Mesgarani, M. Slaney, and S. 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.
- [5] 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.
- [6] X. Domont, M. Heckmann, F. Joubin, and C. Goerick, "Hierarchical spectro-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.
- [7] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Tech. Rep., Apple Computer Co., 1993, Technical report #35.
- [8] P.O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.
- [9] R. Leonard, "A Database for Speaker-independent Digit Recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, 1984, vol. 9.
- [10] 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.
- [11] G. Hirsch, "FaNTFiltering and Noise Adding Tool," Tech. Rep., Niederrhein University of Applied Sciences, <http://dnt-kr.hnsr.de/download.html>.
- [12] 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.
- [13] M. Heckmann, *Adaptive Datenfusion für die Audio-visuelle Spracherkennung*, Shaker, 2003.
- [14] C.P. Chen, K. Filali, and J.A. Bilmes, "Frontend Post-Processing and Backend Model Enhancement on the Aurora 2.0/3.0 Databases," in *Int. Conf. on Spoken Lang. Proc.* 2002, ISCA.