Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Supervised vs. Unsupervised Learning of Spectro Temporal Speech Features

Martin Ernst Heckmann

2010

Preprint:

This is an accepted article published in Statistical And Perceptual Audition (SAPA). The final authenticated version is available online at: https://doi.org/[DOI not available]

Supervised vs. Unsupervised Learning of Spectro Temporal Speech Features

Martin Heckmann

Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany
martin.heckmann@honda-ri.de

Abstract

To overcome limitations of purely spectral speech features we previously introduced Hierarchical Spectro-Temporal (HIST) features. We could show that a combination of HIST and standard features does reduce recognition errors in clean and in noise. The HIST features consist of two hierarchical layers where the corresponding filter functions are learned in a data driven way. In this paper we investigate how different learning methodologies applied to the learning of the filters on the second layer influence the performance. We compare Non-negative Matrix Factorization (NMF), Non-negative Sparse Coding (NNSC), and Weight Coding (WC) on a noisy digit recognition task. NMF and NNSC are unsupervised learning algorithms whereas WC also includes class specific information in the learning process. Additionally we investigate how a mismatch between the database used for learning the features and the one employed for training and testing the recognition system influences the performance.

Index Terms: Spectro-temporal, NMF, NNSC, WC, robust speech recognition, auditory

1. Introduction

It is well known that the dynamic aspects of speech play a crucial role in its understanding [1]. Nevertheless common speech features as Mel Ceptstral Coefficients (MFCCs) [2] and RelAtive SpecTrAl Perceptual Linear Predictive (RASTA-PLP) features [3] capture only stationary spectral information. Dynamic information is only later added via the calculation of first and second order derivatives, also referred to as Delta and Double Delta features.

Yet on the other hand recent findings in neurophysiology have shown that the receptive fields in the mammalian auditory cortex are sensitive to spectro-temporal patterns [4]. Such spectro-temporal receptive fields are potentially better suited to capture formant transitions. Different approaches to make this information also available to automatic speech recognition systems have been proposed recently [5, 6, 7, 8, 9]. When dealing with spectro-temporal features a key issue is the selection of the best suited subset of the huge set of possible spectro-temporal patterns. In [5, 6] a rich set of Gabor features was established and then, based on iterative recognition tests, the optimal subset selected. Another approach proposed in [10] is to subdivide a rich set of Gabor features, interpret them as different streams, and train a recognition system combining the results of these different streams in a multi-stream recognition framework [11].

To deal with the dimensionality problem we previously presented Hierarchical Spectro-Temporal (HIST) features [12, 13]. They consist of two layers, the first capturing local spectro-temporal variations and the second integrating them into larger receptive fields (compare Fig. 1). This layout was inspired by a recently proposed system for visual object recognition [14]. At both layers the receptive fields are learned in a data-driven unsupervised way. On the first layer we apply ICA (Independent Component Analysis) and in the second layer we applied so far Non-Negative Sparse Coding (NNSC). Finally we use a Principal Component Analysis (PCA) to orthogonalize the features and further reduce their dimensionality followed by a Hidden Markov Model (HMM) for the recognition.

In this paper we will investigate alternative learning algorithms for the receptive fields on the second layer of our feature hierarchy. We will compare the two unsupervised approaches Non-negative Matrix Factorization (NMF) and Non-Negative Sparse Coding (NNSC) as well as Weight Coding (WC) an extension of NNSC which takes class specific information during the learning into account.

In this context we will also investigate how the performance of the features is influenced if fundamentally different datasets are used to learn the features and to train and test the subsequent recognition system.

The rest of the paper is organized as follows. In Section 2 we will briefly describe our Hierarchical Spectro-Temporal (HIST) feature extraction framework. This is followed by a description of the different algorithms we apply to learn the features on the second layer in Section 3. The experimental conditions and recognition results will be presented in Section 4. A conclusion and a discussion in Section 5 will close the paper.

2. Hierarchical Spectro Temporal Features

The main building blocks of our hierarchical feature extraction framework are a preprocessing to enhance the formant structure in the spectrograms, a calculation of local and combination features, and a Principal Component Analysis (PCA) to reduce the feature dimension (compare Fig. 1).

2.1. Preprocessing

We apply a Gammatone filter bank to transform the speech signal sampled at 16 kHz into the frequency domain. The filter bank has 128 channels ranging from 80 Hz to 8 kHz and follows the implementation of [15]. From this we obtain spectrograms by rectification and low-pass filtering of the filter bank response. The sampling rate of the spectrograms is then reduced to 400 Hz(compare Fig. 2 a).

An enhancement of the formant structure in the signal is obtained by a pre-emphasis of $+6\,\mathrm{dB/oct.}$ and a subsequent

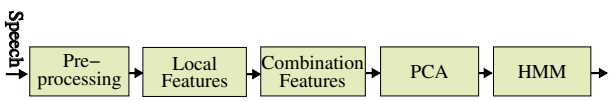
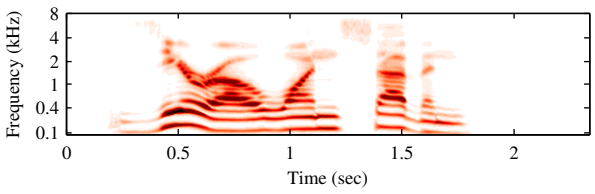
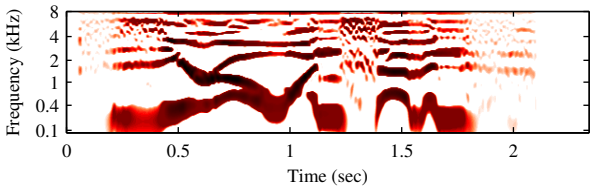


Figure 1: Overview of the feature extraction framework.



(a) Spectrogram



(b) Enhanced formants

Figure 2: Spectrogram of the digit sequence "zero four seven" uttered by a male speaker before (a) and after the formant enhancement (b).

filtering along the frequency axis with a Mexican Hat filter. The last step removes the harmonic structure of the spectrograms and forms peaks at the formant locations (compare Fig. 2 b and see [13] for details).

2.2. First layer: Extraction of local features

In the first layer $Q^{(1)}$ of our hierarchical feature extraction framework local features are extracted via a 2D filtering with a set of $l = 1 \dots n_1$ receptive fields $\boldsymbol{w}_l^{(1)}$, taking the absolute value of the response:

$$q_l^{(1)}(t,f) = \left| \left(\boldsymbol{S} * \boldsymbol{w}_l^{(1)} \right)(t,f) \right|, \tag{1}$$

where the responses $\boldsymbol{q}_l^{(1)}$ of each neuron has the same size as the input spectrogram $\boldsymbol{S}$. The filtering, i. e. convolution, operation is depicted by $*$.

These $n_1 = 8$ receptive fields are 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 $q_l^{(1)}(t,f)$ of the l-th neuron reveals how close a local patch of $\boldsymbol{S}$ centered in (t, f) is to the pattern l. For each local patch only the highest correlated patterns are of interest. Therefore, we perform a Winner-Take-Most (WTM) competition which inhibites the response of the less active neurons at the position (t, f):

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

where $M(t,f) = \max_k q_k^{(1)}(t,f)$ 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 [14].

Furthermore, a nonlinear transformation including a threshold θ_1 is applied on all the $r_l^{(1)}(t,f)$:

$$s_l^{(1)}(t,f) = H(r_l^{(1)}(t,f) - \theta_1), \tag{3}$$

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

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

$$c_l^{(1)}(t,f) = \left(\boldsymbol{s}_l^{(1)} * \boldsymbol{g}_1 \right)(4t, 4f), \tag{4}$$

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

2.3. Second layer: Extraction of combination features

Each of the $k = 1 \dots n_2$ combination patterns on the second layer $Q^{(2)}$ of our hierarchy is composed of n_1 receptive fields $\boldsymbol{w}_{l,k}^{(2)}$, i. e. one for each of the neurons in the previous layer. The coefficients of these receptive fields are non-negative and span all frequency channels. Similarly to (1) the activity $q_k^{(2)}(t)$ of the k-th neuron at time t is given by:

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

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 $\boldsymbol{w}_{l,k}^{(2)}$ are only shifted in the time direction. Note that the absolute value is not required in (5) as both the $\boldsymbol{c}_l^{(1)}$ and the $\boldsymbol{w}_{l,k}^{(2)}$ are non-negative.

This yields $n_2 = 50$ dimensional features $q_k^{(2)}(t)$ at a feature rate of 100 Hz. Delta and double-delta features are computed using a 9th order FIR lowpass and bandpass, respectively. When combining the features $q_k^{(2)}(t)$ with their deltas we obtain an $n = 150$ dimensional vector. This is reduced to a 39 dimensional feature vector $\boldsymbol{x}$ via PCA.

The different algorithms used for the learning of the $n_2 = 50$ receptive fields $\boldsymbol{w}_k^{(2)}$ will be detailed in section 3.

3. Combination Feature Learning

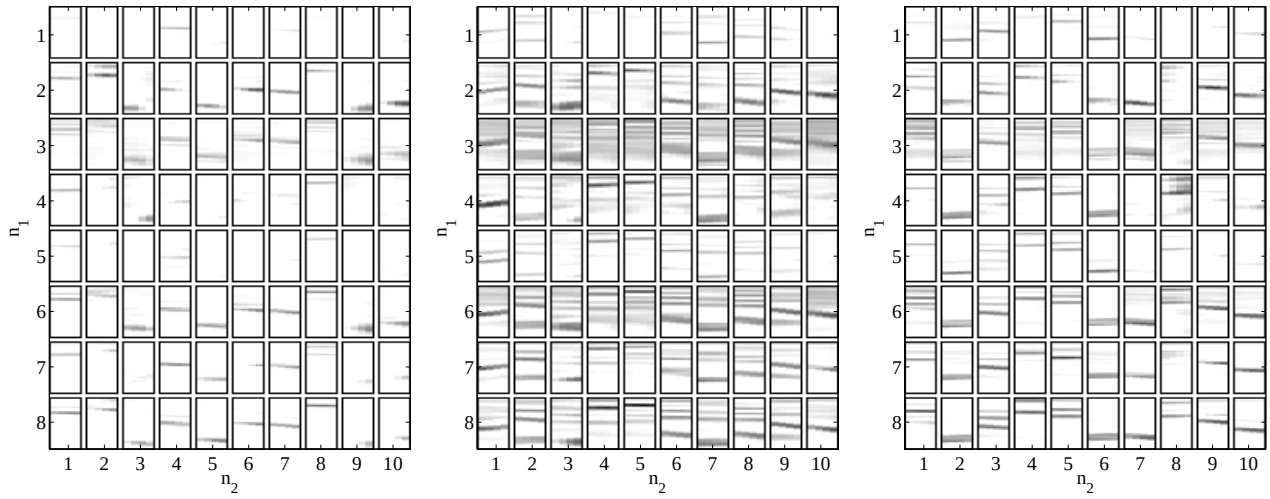
The features on the first, i. e. $Q^{(1)}$, layer of our feature hierarchy are local and rather unspecific. To create more specific features integrating larger regions in time and frequency we investigate different approaches to learn the receptive fields $\boldsymbol{w}_k^{(2)}$ on the second layer.

3.1. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) has been introduced in image processing as a representation which learns in an unsupervised way the relevant parts of an object [16] and has been used in audio processing before [17, 18, 19]. This contrasts to approaches as PCA which represents the images as a whole, i. e. holistically. The main assumption in NMF is that the input data to be represented, the basis functions of the factorization, and the weights at which the basis functions are applied are all positive. As our input data, the spectrograms, are positive we can directly apply it. More precisely we cut out patches $\boldsymbol{P}$ of length $\Delta = 40$ ms of the first layer activations $\boldsymbol{c}_l^{(1)}$. From these patches we learn $n_2 = 50$ combination features by minimizing the following cost function [16]:

$$E = \sum_i \|\boldsymbol{P}_i - \sum_{k=1}^{n_2} \alpha_{k,i} \boldsymbol{w}_k^{(2)}\|^2, \tag{6}$$

where $\boldsymbol{P}_i$ is a tensor representing the n_1 layers of the i-th patch, the $\boldsymbol{w}_k^{(2)}$ are n_2 non-negative tensors each of them containing the n_1 receptive fields $\boldsymbol{w}_{l,k}^{(2)}$, and the $\alpha_{k,i}$ are nonnegative re-



(a) NMF (b) NNSC (c) WC

Figure 3: Selection of 10 receptive fields $\boldsymbol{w}^{(2)}$ resulting from a learning using (a) NMF, (b) NNSC, and (c) WC. The size of the features is 32 channels $\times$ 40 ms $\times n_1$, where $n_1 = 8$ is given by the dimensionality of the $Q^{(1)}$ layer.

construction factors.

In Fig. 3 a a subset of the resulting features is shown. As one can see the features are quite localized.

3.2. Non-negative Sparse Coding

An extension to NMF is Non-negative Sparse Coding (NNSC) which, in addition to the constraints underlying NMF, also puts a constraint on the coefficients to obtain an efficient use of the basis [20]. This is obtained via a so called sparsity term λ which favors reconstructions of $\boldsymbol{P}$ with a sparse usage of the basis $\boldsymbol{w}^{(2)}$ via a minimization of the weights $\boldsymbol{\alpha}$:

$$E = \sum_i \|\boldsymbol{P}_i - \sum_{k=1}^{n_2} \alpha_{k,i} \boldsymbol{w}_k^{(2)}\|^2 + \lambda \sum_i \sum_{k=1}^{n_2} |\alpha_{k,i}| \; . \quad (7)$$

Consequently, if multiple possible reconstructions exist, those are preferred with the more sparse usage of the basis which leads to more complex basis functions. This can also be seen when comparing Fig. 3 a and b. The basis function of the NNSC represent notably more complex combinations of frequency bands than the NMF.

3.3. Weight Coding

The two learning algorithms presented so far where completely unsupervised, i. e. they are not using any class specific information. However, basic functions which mainly capture the information characteristic for a specific class could be beneficial. This can be obtained by introducing another term κ in the cost function (7) of NNSC which penalizes correlations between projections of patches $\boldsymbol{P}_i$ and $\boldsymbol{P}_j$ from two different classes with the same basis function $\boldsymbol{w}_2^k$ [21]:

$$E = \sum_i \|\boldsymbol{P}_i - \sum_{k=1}^{n_2} \alpha_{k,i} \boldsymbol{w}_k^{(2)}\|^2 + \lambda \sum_i \sum_{k=1}^{n_2} |\alpha_{k,i}| + \frac{1}{2}\kappa \sum_k \sum_{\substack{i,j \\ q(i)\neq q(j)}} \frac{\boldsymbol{w}_k^{(2)^T} \boldsymbol{P}_i}{n_{q(i)}} \frac{\boldsymbol{w}_k^{(2)^T} \boldsymbol{P}_j}{n_{q(j)}} \; , \quad (8)$$

where $q(i)$ denotes the class label of $\boldsymbol{P}_i$, $n_{q(i)}$ is the number of samples in the class of $\boldsymbol{P}_i$, and T denotes the transpose operator. The resulting basis is depicted in Fig. 3 c. As can be seen the additional class specific term seems to counteract the sparsity constraint to some extent as the resulting features are of less global character as those resulting from NNSC and at the same time not as local as those resulting from NMF.

4. Results

We compare the performance of the different combination feature learning approaches in a noisy digit recognition task. For doing so we added to TIDigits [22], a database for speaker independent continuous digit recognition, white noise, noise recorded in a factory and in a car and babble noise, all taken from the Noisex database [23] at Signal to Noise Ratios (SNRs) ranging from $-5\,\mathrm{dB}\ldots$ inf, i. e. we also kept the clean signal. The HMMs were trained with HTK [24] using whole word HMMs containing 16 states without skip transitions and a mixture of 3 Gaussians with a diagonal covariance matrix per state.

For the Weight Coding learning scheme we decided to use the phoneme classes as underlying classes. As TIDigits does not contain phonetic transcriptions we used TIMIT [25] instead to train the receptive fields of the $Q^{(1)}$ and $Q^{(2)}$ layer of our feature hierarchy as well as the final PCA. We did so not only for WC but also for NMF and NNSC. We identified 21 phonemes necessary to cover the digit sequences in TIDigits and randomly extracted for each of these phonemes 3000 segments of length $40\,\mathrm{ms}$ from TIMIT. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences [25]. Silences and pauses were not included as phonemic categories. We set the factor of the WTM competition $\gamma_1 = 0.7$, the thresholding constant $\theta_1 = 0.25$, the sparsity factor $\lambda = 0.05$ and the weight factor $\kappa = 0.8$. These values were determined heuristically. As benchmark we also extracted RASTA-PLP features [3].

As one can see from Fig. 4 for high SNR values RASTA-PLP features perform better than the HIST features with all

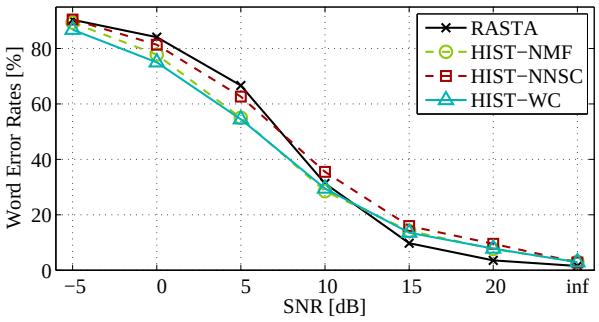


Figure 4: Word error rates (WERs) for HIST and RASTA-PLP features when factory noise was added at SNR levels ranging from $-5\,\mathrm{dB}\ldots$ inf.

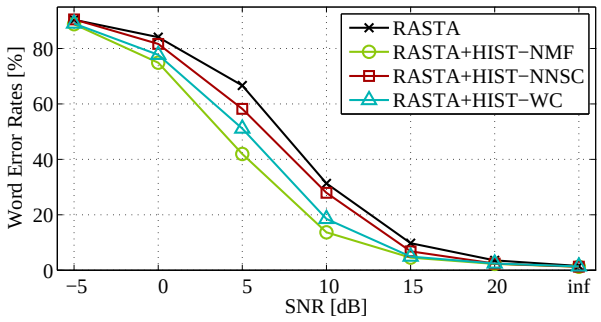


Figure 5: Word error rates (WERs) for the combination of HIST and RASTA-PLP features as well as RASTA-PLP features alone when factory noise was added at SNR levels ranging from $-5\,\mathrm{dB}\ldots$ inf.

	white	factory	babble	car
RASTA-PLP	43.1	41.0	35.0	19.5
HIST-NMF	41.2	39.4	55.7	16.3
HIST-NNSC	44.4	42.6	64.7	16.3
HIST-WC	40.2	38.6	58.4	16.4
RASTA-PLP+HIST-NMF	32.9	32.5	49.5	11.4
RASTA-PLP+HIST-NNSC	38.0	38.4	59.7	12.8
RASTA-PLP+HIST-WC	35.9	35.0	58.4	12.4
RASTA-PLP+HIST-NMF$_{\mathrm{TI}}$	27.9	30.3	49.0	10.6
RASTA-PLP+HIST-NNSC$_{\mathrm{TI}}$	30.1	31.4	44.8	11.6

Table 1: Average word error rates for the different feature types when the specified noise types at SNR values ranging from $-5\,\mathrm{dB}\ldots$ inf were added.

types of combination features. On the other hand for low SNR values the HIST features obtain similar or better results than the RASTA-PLP features. Yet the difference between the different types of HIST features is small. In Table 1 we also depicted the values for the different noise types averaged over all SNR levels.

The behavior we observe suggests to investigate a combination of the HIST and RASTA-PLP features. In these cases the combination was obtained via feature concatenation, i. e. we concatenated the 39 dimensional HIST features and the 45 dimensional RASTA-PLP features to a 84 dimensional feature vector. The results of this feature concatenation can be seen in Fig. 5 and Table 1.

One can see that the combination of HIST and RASTA-PLP features improves results, especially for medium and high SNR values. To better asses this we also calculated the relative improvements of the feature combination compared to RASTA-PLP features alone (compare Fig. 6 and Table 2).

This reveals that the combination of HIST and RASTA-PLP features independent of the learning algorithm improves results for all noise types and SNR levels with the exception of babble noise. We have seen this unfavorable behavior of the HIST features in babble noise already previously [13]. Via additional experiments we concluded that the reason for this is the very high sensitivity of the HIST features to speech. The preprocessing strongly enhances the speech structures present in babble noise and leads to a very significant amount of word insertions. This causes the very unfavorable recognition results. As we have shown in [13] this can be remedied by inserting also babble noise in the training phase. In this case the subsequent HMMs

	white	factory	babble	car
RASTA-PLP+HIST-NMF	29.6	30.8	-145.7	41.0
RASTA-PLP+HIST-NNSC	10.3	15.3	-292.4	39.4
RASTA-PLP+HIST-WC	19.2	23.7	-268.0	36.0
RASTA-PLP+HIST-NMF$_{\mathrm{TI}}$	35.9	33.3	-123.7	41.4
RASTA-PLP+HIST-NNSC$_{\mathrm{TI}}$	34.0	32.1	-85.1	40.7

Table 2: Average relative improvement of the combination of HIST and RASTA-PLP features when the specified noise types at SNR values ranging from $-5\,\mathrm{dB}\ldots$ inf were added.

learn to discriminate between real speech segments and babble noise.

When we analyze the different learning algorithms more closely we see that NMF shows the best performance. NNSC and WC perform very similar to NMF for medium to high SNR values but show clear inferior behavior at low SNR values. Thereby the performance of WC lies in between those of NMF and NNSC.

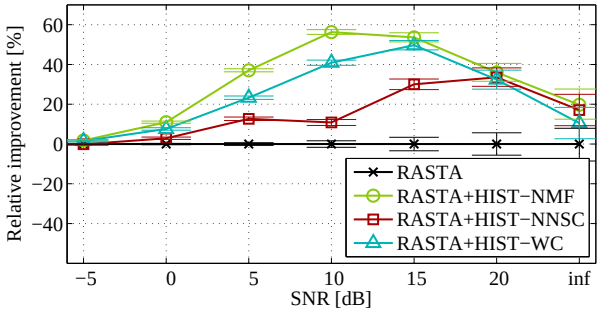


Figure 6: Relative improvements compared to RASTA-PLP features when factory noise was added to the test set. The bars indicate the 95% confidence intervals calculated according to [26].

We saw already in previous experiments such substantial improvements from the combination of HIST features and RASTA-PLP features [27, 13]. Yet in those experiments only NNSC was deployed and the features were learned on the training set of TIDigits, the database on which also the recognition tests are performed and not as in this case on TIMIT which dif-

fers substantially from TIDigits. Recall that TIDigits consists of continuously uttered digit sequences whereas TIMIT contains phonetically rich sentences which do not contain numbers.

In a second experiment we investigated to what extent the database used in the learning of the features influences the performance. Hence we learned for NMF and NNSC both the $Q^{(1)}$ and the $Q^{(2)}$ layer on the TIDigits database. As TIDigits does not contain phonetic labels we could not include WC in this test. The relative improvements of the combination of HIST and RASTA-PLP features for this setup are given in Table 2 and Fig. 7 (the subscript TI indicates that the training of the receptive fields was performed on TIDigits).

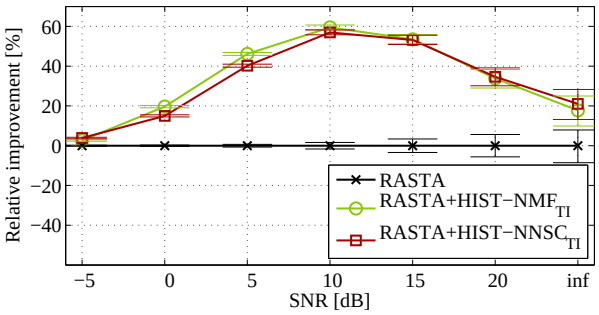


Figure 7: Relative improvements of the features trained on TIDigits compared to RASTA-PLP features when factory noise was added to the test set. The bars indicate the 95% confidence intervals calculated according to [26].

As one can see in case of the NMF results obtained when learning the features on TIDigits are very similar to those obtained when learning them on TIMIT. Overall the results in the case where the database used for learning the features and performing the recognition experiments, i. e. in both cases TIDigits, are slightly better. However, the performance of NNSC improved significantly in this matched learning condition. In this case the difference between NMF and NNSC is only small.

5. Conclusion

In previous experiments we could already show that the unsupervised learning in our proposed hierarchical spectro-temporal speech features is able to extract from the very high dimensional space of spectro-temporal patterns information not captured by conventional spectral features and that this information is beneficial to improve recognition results [12, 13]. Here we investigated how alternative learning algorithms applied on the second layer of our hierarchy influence the performance and if the features learned are specific to a single database or capture general speech properties.

The results showed that the impact of changing the learning algorithm is rather low for medium to high SNR levels. Including class specific information in the learning of the features as in the Weight Coding did not yield better recognition scores than NMF or NNSC. For low SNR levels NMF outperformed the other two approaches. It seems that the more local receptive fields resulting from NMF are better suited in these cases. The class specific information available to WC seems to be able to counterbalance to some extent the unfavorable effect of the sparsity constraint used in NNSC. Hence, we will investigate in the future a learning approach which uses the class specific information but without the sparsity constraint.

Regarding the question of the database specificity of the learned features we could show that the performance deteriorated only little from the case when learning of the features as well as training and testing of the recognition system were performed on the same database to the case when we used two different databases. This is true for NMF for all SNR levels and for the other two learning algorithms only for high SNR levels. As the two databases we compared during learning of the features, namely TIDIgits and TIMIT, cover a quite different domain, we conclude that the information captured by the HIST features when using NMF for learning is indeed not database but speech specific. To what extent it is language specific has to be determined in further experiments.

Concerning the rather disappointing results regarding the inclusion of class specific information via the WC algorithm we suppose that our approach of using only 21 phoneme classes to capture the information relying in the formant transitions is not optimal. A modeling based on diphones or triphones seems more promising. Yet it might also be an inherent problem of the WC algorithm as previous experiments in a visual object recognition task could also not identify a clear benefit from the inclusion of the class specific term in the learning [21].

6. Acknowledgments

We want to thank Stephan Hasler and Dr. Heiko Wersing for supporting us with details on the different combination layer learning algorithms and providing us their implementation of the learning algorithms.

7. References

[1] S. Furui, “On the role of spectral transition for speech perception,” *The Journal of the Acoustical Society of America*, vol. 80, no. 4, pp. 1016–1025, 1986.

[2] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 28, no. 4, pp. 357–366, 1980.

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

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

[5] M. Kleinschmidt and D. Gelbart, “Improving word accuracy with Gabor feature extraction,” in *Proc. Int. Conf. on Spoken Language Proc. (ICSLP)*, Denver, CO, 2002, ISCA.

[6] B.T. Meyer and B. Kollmeier, “Complementarity of MFCC, PLP and Gabor features in the presence of speech-intrinsic variabilities,” in *Proc. INTERSPEECH*, Brighton, UK, 2009.

[7] T. Ezzat and T. Poggio, “Discriminative word-spotting using ordered spectro-temporal patch features,” in *Proc. SAPA*, Brisbane, 2008.

[8] S. Y. Zhao and N. Morgan, “Multi-stream spectro-temporal features for robust speech recognition,” in *Proc. INTERSPEECH*, Brisbane, 2008.

[9] F. Valente and H. Hermansky, “Discriminant linear processing of time-frequency plane,” in *Proc. INTERSPEECH*, Pittsburgh, PA, 2006.

[10] S.Y. Zhao, S. Ravuri, and N. Morgan, “Multi-Stream to Many-Stream: Using Spectro-Temporal Features for ASR,” in *Proc. INTERSPEECH*, Brighton, UK, 2009.

[11] H. Bourlard, S. Dupont, F. TCTS, and P. de Mons, “A mew ASR approach based on independent processing andrecombination of partial frequency bands,” in *Proc. Int. Conf. Spoken Language Proc. (ICSLP)*, Philadelphia, PA, 1996.

[12] 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, NV, 2008, pp. 4417–4420.

[13] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *accepted for Speech Communication*, 2010.

[14] 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.

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

[16] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[17] S. Behnke, "Discovering hierarchical speech features using convolutional non-negative matrix factorization," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, 2003, vol. 4, pp. 2758–2763.

[18] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[19] Y.C. Cho, S. Choi, and S.Y. Bang, "Non-negative component parts of sound for classification," in *Proc. 3rd IEEE Int. Symposium Signal Proc. and Information Technology, 2003. ISSPIT 2003*, 2003, pp. 633–636.

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

[21] S. Hasler, H. Wersing, and E. Körner, "Combining reconstruction and discrimination with class-specific sparse coding," *Neural Computation*, vol. 19, no. 7, pp. 1897–1918, 2007.

[22] R. Leonard, T.I. Incorporated, and T. Dallas, "A database for speaker-independent digit recognition," in *Int. Conf. Acoustics, Speech, and Signal Proc. (ICASSP)*, San Diego, CA, 1984, vol. 9, IEEE.

[23] 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.

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

[25] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren, *DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM*, Philadelphia, 1993.

[26] J.M. Vilar, "Efficient computation of confidence intervals for word error rates," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, Las Vegas, NV, 2008, pp. 5101–5104, IEEE.

[27] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A closer look on hierarchical spectro-temporal features (HIST)," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008.