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

A Hierarchical Framework for Spectro-Temporal Feature Extraction

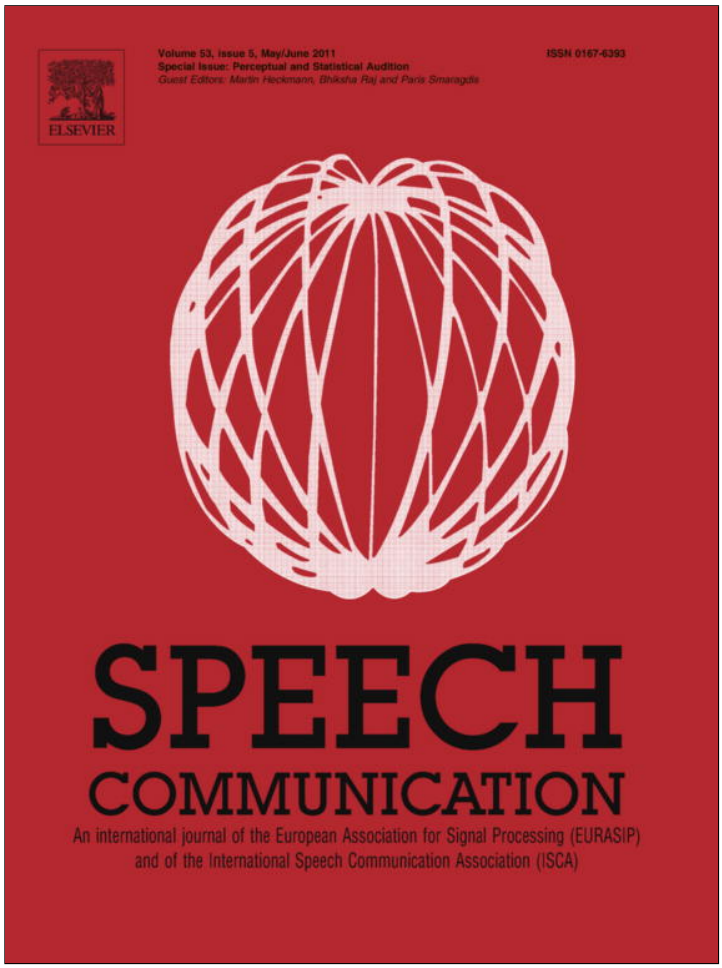
Martin Heckmann, Xavier Domont, Frank Joublin, Christian Goerick

2011

Preprint:

This is an accepted article published in Speech Communication. The final authenticated version is available online at: https://doi.org/[DOI not available]

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 736–752
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A hierarchical framework for spectro-temporal feature extraction

Martin Heckmann [a,*], Xavier Domont [a,b], Frank Joublin [a], Christian Goerick [a]

[a] Honda Research Institute Europe GmbH, D-63073 Offenbach am Main, Germany
[b] Technische Universität Darmstadt, Control Theory and Robotics Lab, D-64283 Darmstadt, Germany

Available online 11 August 2010

Abstract

In this paper we present a hierarchical framework for the extraction of spectro-temporal acoustic features. The design of the features targets higher robustness in dynamic environments. Motivated by the large gap between human and machine performance in such conditions we take inspirations from the organization of the mammalian auditory cortex in the design of our features. This includes the joint processing of spectral and temporal information, the organization in hierarchical layers, competition between coequal features, the use of high-dimensional sparse feature spaces, and the learning of the underlying receptive fields in a data-driven manner. Due to these properties we termed the features as hierarchical spectro-temporal (HIST) features. For the learning of the features at the first layer we use Independent Component Analysis (ICA). At the second layer of our feature hierarchy we apply Non-Negative Sparse Coding (NNSC) to obtain features spanning a larger frequency and time region. We investigate the contribution of the different subparts of this feature extraction process to the overall performance. This includes an analysis of the benefits of the hierarchical processing, the comparison of different feature extraction methods on the first layer, the evaluation of the feature competition, and the investigation of the influence of different receptive field sizes on the second layer. Additionally, we compare our features to MFCC and RASTA-PLP features in a continuous digit recognition task in noise. On a wideband dataset we constructed ourselves based on the Aurora-2 task, as well as on the actual Aurora-2 database. We show that a combination of the proposed HIST features and RASTA-PLP features yields significant improvements and that the proposed features carry complementary information to RASTA-PLP and MFCC features.
© 2010 Elsevier B.V. All rights reserved.

Keywords: Spectro-temporal; Auditory; Robust speech recognition; Image processing; Learning; Competition; Hierarchical

1. Introduction

Humans have the astonishing ability to preserve stable representations even in a dynamic environment. In contrast to automatic speech recognition systems additional background noise and changes in the transmission channel have only minor effects on human performance (Lippmann, 1997; Sroka and Braida, 2005). Hence better understanding the underlying processes in humans bears the potential to yield better recognition systems.

Unfortunately, our knowledge of the auditory processing in the mammalian brain is still quite limited. The visual system is for example already much better understood (King and Nelken, 2009). This is also expressed in the wealth of corresponding computational models of the visual system. On the other hand, there are several studies highlighting important similarities between the two systems. Sur et al. (1988) showed that newborn ferrets whose retinal nerves were rerouted to the auditory part of the thalamus, sometimes called the gateway to the cortex (Crick, 1984), where later able to respond to visual stimuli via their auditory cortex. This at least demonstrates a high plasticity of these areas during development if not a strong similarity in functional organization. Despite significant differences between auditory and visual processing in the brain common, modality independent processing principles are usually assumed (Read et al., 2002; King and Nelken, 2009). Different authors have shown that at least at the level of the receptive fields in the primary visual and

* Corresponding author. Tel.: +49 69 8901 1755.
E-mail addresses: martin.heckmann@honda-ri.de (M. Heckmann), xavier.domont@rtr.tu-darmstadt.de (X. Domont), frank.joublin@honda-ri.de (F. Joublin), christian.goerick@honda-ri.de (C. Goerick).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.08.006

Author's personal copy

auditory cortices these similarities can be found (Schreiner and Calhoun, 1994; de Charms et al., 1998; Shamma, 2001). Measurements in the primary auditory cortex of different animals revealed its spectro-temporal organization, i.e. the receptive fields are selective to modulations in the time-frequency domain. The corresponding receptive fields have, as in the visual cortex, Gabor-like shapes. The above mentioned findings suggest that modeling principles known from image processing can beneficially be transferred to auditory tasks.

Traditionally, speech features mainly took inspirations from psychoacoustic findings and thereby relied in most cases on independent spectral (Hermansky, 1990; Hermansky and Morgan, 1994; Flynn and Jones, 2008; Haque et al., 2009) or temporal representations (Hermansky and Sharma, 1998).

In recent years also features inspired by above mentioned similarities between visual and auditory processing, i.e. features capable of directly capturing spectro-temporal variations, were developed. In his seminal work Kleinschmidt (2002) introduced the usage of 2 D Gabor features for speech recognition in noise. This was followed by others, also employing Gabor features on similar tasks, including speech vs. non-speech discrimination (Mesgarani et al., 2006; Meyer and Kollmeier, 2008; Sherry and Zhao, 2008). In (Elhilali and Shamma, 2006) a spectro-temporal respresentation based on Gabor filters was used for source separation. Ezzat et al. (2007) randomly selected spectro-temporal patches of the target word from the training set and then used these as features for keyword spotting.

The framework for the extraction of spectro-temporal speech features we present here takes many inspirations from the visual object recognition system of Wersing and Körner (2003) and is an extension of our previous work (Domont et al., 2007, 2008). In contrast to the previously mentioned approaches and other models in the literature we integrated additional processing principles which are also inspired by the mammalian sensory cortex. The processing in the sensory cortices seems to be organized in a hierarchical fashion. This has been stated for the visual (Hubel and Wiesel, 1965; Felleman and Van Essen, 1991) and auditory cortex (Rauschecker, 1998; Read et al., 2002; Scott et al., 2003). Based on this principle we propose a hierarchical framework consisting of two layers.[1] Features in the first layer extract local information. In the second layer the results of these local features are combined to form more complex features. The hierarchical processing and the construction of more complex and at the same time more specific features leads to a substantial increase in the number of features. A trend also observed in the human brain where approximately 3500 inner hair cells are present in the cochlea and about 100,000,000 neurons in the auditory cortex (Dusan and Rabiner, 2005).

In general, when dealing with spectro-temporal features developing methods for the selection of the relevant features is a key issue. Kleinschmidt (2002) already proposed to use a so-called "feature finding neural network" to select the features yielding the best recognition rates. The unsupervised learning of sparse spectro-temporal representations was proposed in (Klein et al., 2003; Behnke, 2003). Cho and Choi (2005) investigated how such learned representations can be applied to the task of sound classification. We follow this idea in that we learn the receptive fields on both layers of our hierarchy with unsupervised learning rules. Another important property of the mammalian sensory cortices is the competition between coequal features. To model this we implemented a Winner-Take-Most competition between the features on the first layer.

The following sections will describe our framework in more detail and will evaluate its performance in comparison to conventional speech features. In Section 2 we give an overview on our framework. Section 3 describes the preprocessing we apply to the speech signal prior to the feature extraction. The learning of the receptive fields and the extraction of the spectro-temporal features is detailed in Section 4. Section 5 presents recognition results we obtain on a noisy digits task. Based on this task this section also evaluates the contribution of the different elements of our framework to the overall performance. Finally, in Section 6 we discuss the results we obtain and possible improvements.

2. Overview

The key elements of our hierarchical feature extraction framework are depicted in Fig. 1. The first step performs a preprocessing of the speech signal and mainly consists of a transformation into the frequency domain and an enhancement of the formant structure. Based on this we calculate local spectro-temporal features. This step is followed by a competition between these local features. Together these two steps constitute the first layer of our framework. On the second layer the local features are combined to form complex features, spanning larger time and frequency regions. The final steps are an orthogonalization of the features via a Principal Component Analysis (PCA) and recognition of the feature stream with a Hidden Markov Model (HMM) based recognizer. Strictly speaking we do not consider the last two steps as being part of our framework. However, they are necessary to evaluate the feature extraction. Due to their hierarchical organization we termed the features resulting from the proposed framework as hierarchical spectro-temporal (HIST) features.

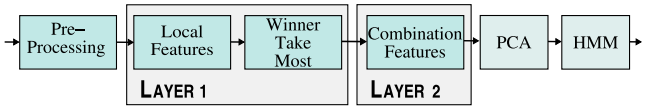


Fig. 1. Overview of the feature extraction process.

[1] Behnke (2003) already suggested a hierarchical speech feature extraction framework but did not evaluate the resulting features in respect to what information they extract and how this could be used for speech processing.

Author's personal copy

3. Preprocessing

As most of the phonetic information is conveyed via the formant variations we perform a preprocessing of the signal after transformation into the spectral domain which aims at enhancing the formant structure.

The process of speech production is commonly modeled via a non-linear volume velocity source followed by a time-varying linear filter and radiation components (Fant, 1970). Hence, the speech signal we hear is the overlay of the excitation signal at the glottis, the time-varying resonance frequencies of the vocal tract, the radiation components of the mouth and lips, and influences of the room. The preprocessing we present in the following was mainly developed in (Gläser et al., 2010). It aims at compensating for the effects of the excitation signal and the radiation components and thereby focuses on the resonance frequencies of the vocal tract, commonly referred to as formants. We do not explicitly model the room effects or the excitation signal.

The spectral tilt introduced by excitation and radiation can be corrected via a pre-emphasis. Additionally, for voiced sounds the glottis converts the steady airflow produced by the lungs into a quasi-periodic train of flow pulses by which the transfer function of the vocal tract is sampled at multiples of the fundamental frequency. Consequently, spectrograms feature spectral peaks at the harmonics rather than the vocal tract resonance frequencies. This effect will be compensated for by a smoothing along the frequency axis.

3.1. Gammatone filterbank

We transform the speech signal into the spectro-temporal domain via the Patterson–Holdsworth auditory filterbank (Patterson et al., 1992). This filterbank is based on neurophysiological findings on the human auditory system and models the peripheral processing as carried out by the cochlea, where sound is transformed into spatio-temporal response patterns on the auditory nerve. It is implemented as a set of linear Gammatone filters, each of them tuned to a different frequency range. Hence, non-linear effects as suppression and level-dependent tuning curves are not modeled. The filterbank we use is composed of 128 filters covering the frequency range from 80 Hz to 8 kHz and follows the implementation suggested by Slaney (1993). Subsequently, the spectral envelope is calculated via rectification and low-pass filtering. Fig. 2(a) exemplary shows the envelope of the filter responses to a sequence of digits from the TIDigits database (Leonard et al., 1984).

3.2. Pre-emphasis

The source signal of voiced sounds is produced at the glottis. Fant (1979) suggested to use a second-order low-pass filter for the approximation of the glottal flow spectrum. This adequately approximates the most common phonation types, modal and creaky phonation (Childers and Lee, 1991). Thus, voiced excitation changes the spectral characteristic by −12 dB/oct.

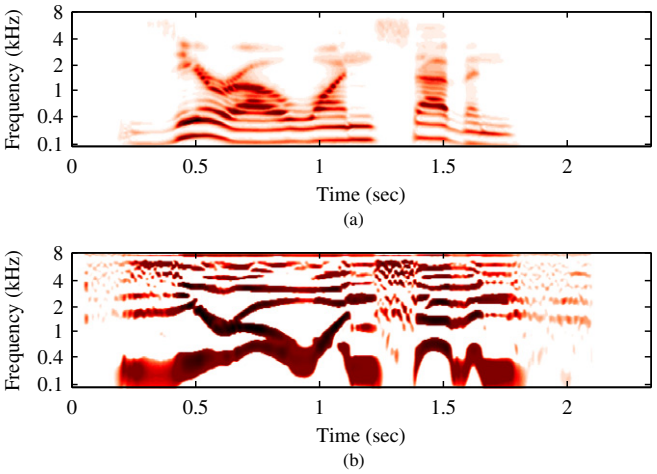


Fig. 2. In (a) the original spectrogram for the digit sequence "zero-one-seven" spoken by a male speaker is shown. The spectrogram after the application of the preprocessing for enhancing the formant structure is shown in (b).

Following (Stevens, 2000) the radiation components, i.e. the lip impedance, can be modeled via a first-order high-pass filter yielding a spectral change of +6 dB/oct. Overall this means that a pre-emphasis via amplification of frequency magnitudes by +6 dB/oct adequately eliminates the spectral influence of excitation and radiation.

3.3. Spectral filtering

After the above mentioned pre-emphasis we enhance the formant structure in the spectrogram by smoothing along the frequency axis following the same spirit as (Baer et al., 1993). We obtain this by a filtering with channel-dependent Difference-of-Gaussians (DoG) operators with standard deviations of the negative Gaussian components being twice as large as that of the corresponding positive ones:

$$\mathrm{DoG}_k(f) = \frac{1}{\sqrt{2\pi}}\left(\exp\left(-\frac{(f-fc_k)^2}{2\sigma^2_{\mathrm{DoG}_k}}\right) - \frac{1}{2}\exp\left(-\frac{(f-fc_k)^2}{8\sigma^2_{\mathrm{DoG}_k}}\right)\right). \quad (1)$$

Here, $\mathbf{DoG}_k$ is the DoG operator of channel k featuring a center frequency fc_k. As standard deviations σ_{DoG_k} we use 70 Hz for filter channels with center frequencies in the range from 80 Hz to 5 kHz and 400 Hz filter channels with center frequencies in the range from 5 kHz to 8 kHz. We discretized the DoGs by sampling them at the logarithmically arranged Gammatone filterbank's channel center frequencies and normalized the resulting $\mathbf{DoG}_k$ vectors. Fig. 3 exemplarily depicts the DoGs of 8 filter channels. The smoothed spectrogram $\boldsymbol{S}_{\mathrm{Smooth}}$ is obtained via multiplying the matrix

Author's personal copy

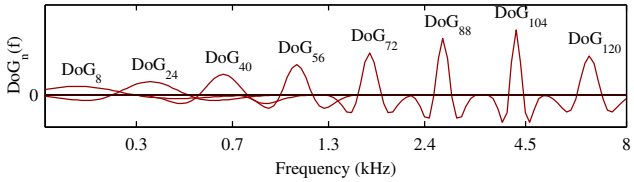


Fig. 3. The DoG operators of the filter channels used to enhance the formant structure in spectrograms vary in bandwidth and frequency resolution dependent on the channels' center frequencies. Here, the DoG operators of 8 exemplarily chosen filter channels are shown.

DoG containing in its rows the vectors $\mathbf{DoG}_k^T$ with the spectrogram following the pre-emphasis $\boldsymbol{S}_{\text{Pre-emphasis}}$:

$$\boldsymbol{S}_{\text{Smooth}} = \mathbf{DoG} \cdot \boldsymbol{S}_{\text{Pre-emphasis}}. \tag{2}$$

The smoothing with the DoG operator results in a suppression of the responses from neighboring frequency channels and thereby sharpens the peaks at the formants. We add a non-linear component to this suppression by suppressing all negative values. Additionally we apply a fourth root compression on the resulting signal to approximate the non-linear loudness perception:

$$\boldsymbol{S}_{\text{Compressed}} = \sqrt[4]{\max(\boldsymbol{S}_{\text{Smooth}}, 0)}. \tag{3}$$

In a final step we filter the spectrograms along the time direction with a low-pass filter having a cut-off frequency of 100 Hz and then perform a downsampling to 400 Hz. Fig. 2(b) continues the example depicted in (a) by showing the corresponding resulting spectrogram $\boldsymbol{S}$.

4. Hierarchical spectro-temporal features

The hierarchical feature extraction framework we present here consists of two layers. The first layer extracts local features, covering small regions in the spectro-temporal domain. In the second layer these local features are integrated to more complex features integrating information over longer time spans and frequency ranges (e.g. 40 ms × 8 kHz). Structuring it in this way was largely inspired by the visual object recognition system of Wersing and Körner (2003). Such hierarchical feature extraction schemes have shown to be beneficial in visual object recognition (Fukushima, 1980; Riesenhuber and Poggio, 1999; Fergus et al., 2003). In the following we will demonstrate how these approaches can be transferred to the auditory domain.

4.1. Extraction of local features

On the first layer $Q^{(1)}$ we extract features via a 2 D filtering of the preprocessed spectrograms $\boldsymbol{S}$ with a set $n^{(1)}$ of receptive fields $\boldsymbol{w}_l^{(1)}$. After filtering we only keep the absolute value of the response:

$$q_l^{(1)}(t,f) = \left| \left(\boldsymbol{S} * \boldsymbol{w}_l^{(1)} \right)(t,f) \right|, \tag{4}$$

The indices t and f indicate time and frequency. As a consequence of the filtering border effects occur when the receptive fields only partially overlap with the input spectrogram. We decided to keep the size of the $n^{(1)}$ spectrograms $\boldsymbol{q}_l^{(1)}$ after the filtering identical to the spectrogram $\boldsymbol{S}$ at the input and use an overlay-save implementation for the filtering in the time direction. This yields small border effects in the frequency direction and in most cases negligible effects in the time direction for the first and last block.

The filtering with the receptive fields, i.e. the filter kernels, is identical to a correlation between the spectrogram at the input and the frequency and time inverted receptive fields. Hence, the responses in the output spectrograms indicate the similarity of the spectro-temporal patches in the input spectrogram to the (time and frequency inverted) receptive fields.

Neurobiological experiments show that the processing in the primary visual cortex aims at reducing the redundancy of the stimuli. When applying corresponding unsupervised learning methods on natural images one obtains receptive fields similar in shape to Gabor functions (Olshausen et al., 1996; van Hateren and Ruderman, 1998).

We therefore decided to learn the filters of the first layer of our feature hierarchy with Independent Component Analysis (ICA) (Comon, 1994). For the learning we used 3500 randomly selected patches of size 50 ms × 20 channels, i.e. 20 × 20 pixels, taken from the training set of the TIDigits database (Leonard et al., 1984). We applied the FastICA fixed-point algorithm (Hyvärinen, 1999) for the calculation of the ICA. A tanh non-linearity was used, i.e. the negentropy of the data is approximated by the contrast function $G(u) = 1/a \log\cosh(au)$, with the parameter a in the range 1–2. For $n^{(1)} = 8$ the resulting receptive fields, depicted in Fig. 4, are rather difficult to interpret. They show some preferences for steady formants and upward and downward transitions. However, overall they are not very specific and only vaguely resemble Gabor functions.

Additionally, we also investigated the usage of genuine Gabor functions. The Gabor functions were not learned but selected manually. During this selection process we investigated different sets of filters. Thereby, we set the

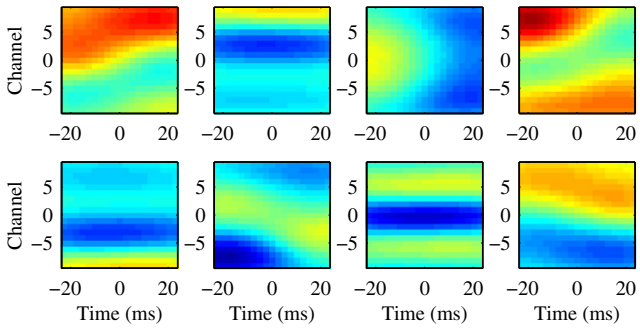


Fig. 4. Visualization of the receptive fields used for the extraction of local features learned via ICA. The frequency axis is given in channels as on the non-linear frequency grid the actual extend in frequency depends on the position on the frequency axis of the receptive field.

Author's personal copy

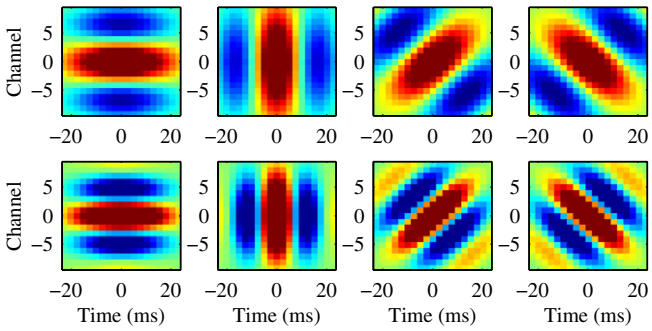


Fig. 5. Visualization of the Gabor based receptive fields used for the extraction of local features. The frequency axis is given in channels as on the non-linear frequency grid the actual extend in frequency depends on the position on the frequency axis of the receptive field.

modulation frequencies and orientations to resemble those of the learned ICA filters and to what one would expect to be a relevant feature given the widths and orientations of the formants in the enhanced spectrograms (see Fig. 5 for an example).

A discrete two dimensional complex Gabor function is a complex sinusoidal carrier $s(t,f)$ modulated with a Gaussian envelope $g(t,f)$. Important properties of Gabor functions are the shape of the Gaussian envelope, controlled via the variances σ_{G_t} and σ_{G_f}, and the frequency and orientation of the carrier, governed by the radian frequencies ω_t and ω_f. Further parameters are the center of mass in time t_0 and frequency f_0. The complex carrier $s(t,f)$ is defined as

$$s(t,f) = \exp\left[i\omega_t(t-t_0) + i\omega_f(f-f_0)\right] \tag{5}$$

and the Gaussian envelope $g(t,f)$ as

$$g(t,f) = \frac{1}{2\pi\sigma_{G_t}\sigma_{G_f}} \cdot \exp\left[\frac{-(t-t_0)_r^2}{2\sigma_{G_t}^2} + \frac{-(f-f_0)_r^2}{2\sigma_{G_f}^2}\right], \tag{6}$$

where the subscript r denotes the rotated coordinates

$$\begin{aligned}(t-t_0)_r &= (t-t_0)\cos\theta + (f-f_0)\sin\theta,\\ (f-f_0)_r &= -(t-t_0)\sin\theta + (f-f_0)\cos\theta\end{aligned} \tag{7}$$

with $\theta = \tan^{-1}(\omega_f/\omega_t)$, if $\omega_f \neq 0$, and $\theta = sign(\omega_f)\pi/2$ otherwise. A modulation frequency $\omega_t = 0$ yields purely spectral filters ($\theta = \pi/2$), a modulation frequency $\omega_f = 0$ purely temporal filters ($\theta = 0$), and settings in between spectro-temporal filters. The real part of the complex Gabor function yields an even filter and the imaginary part an odd filter. For our implementation we used the 8 odd filters depicted in Fig. 5.

4.2. Competition between local features

Ideally the neurons in the previous processing step will only respond to spectro-temporal patterns in the spectrogram matching their receptive field (e.g. an upward moving formant). However, the neurons are not sufficiently selective and most of the time several neurons respond with similar strength to a given pattern. To counterbalance this we introduce a competition mechanism between the different neurons similar to Wersing and Körner (2003). We independently apply a Winner-Take-Most (WTM) competition between the activities $q_l^{(1)}(t,f)$ of the l neurons for all points (t,f) in the spectrogram:

$$r_l^{(1)}(t,f) = \begin{cases} 0 & \text{if } \frac{q_l^{(1)}(t,f)}{M(t,f)} < \gamma^{(1)} \\ & \text{or } M(t,f) = 0 \\ \frac{q_l^{(1)}(t,f) - \gamma^{(1)} M(t,f)}{1-\gamma^{(1)}} & \text{else,} \end{cases} \tag{8}$$

$M(t,f) = \max_k q_k^{(1)}(t,f)$ is the maximal value at position (t,f) over the $n^{(1)}$ neurons and $0 \leqslant \gamma^{(1)} \leqslant 1$ is a parameter controlling the strength of the competition. By suppressing less active neurons the selectivity of the feature extraction is enhanced, a process commonly found in the auditory system (Young, 2008).

Furthermore, a non-linear transformation including a threshold $\vartheta^{(1)}$ is applied on all the $r_l^{(1)}(t,f)$:

$$s_l^{(1)}(t,f) = H(r_l^{(1)}(t,f) - \vartheta^{(1)}), \tag{9}$$

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

To introduce robustness against small fluctuations in location and amplitude as well as further sharpen the contrast between the activities we perform a pooling followed by a non-linear transformation on all the $s_l^{(1)}(t,f)$. The pooling consist of a smoothing with a 2D Gaussian filter $\boldsymbol{g}^{(1)}$ and reduction of the resolution of the activations $\boldsymbol{s}_l(t,f)$ by a factor of four in both frequency and time dimension. As non-linear transformation we use a tanh function:

$$c_l^{(1)}(t,f) = \tanh\left(\boldsymbol{s}_l^{(1)} * \boldsymbol{g}^{(1)}\right)(4t, 4f). \tag{10}$$

This yields 32 frequency channels and a sampling rate of 100 Hz. In the visual object recognition system of (Wersing and Körner, 2003) a binarization was performed after the tanh non-linearity. This results in further robustness against fluctuations. As prerequisite it is required that the dynamics of the input data does not vary too much. For images this can be guaranteed via a normalization of each

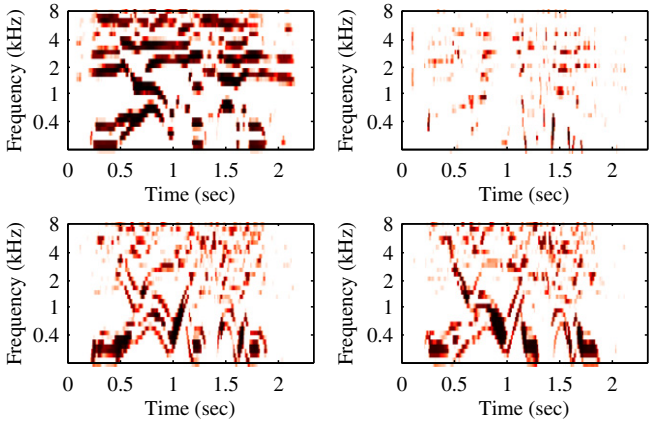


Fig. 6. The results of the first feature extraction layer obtained with the example from Fig. 2 are visualized. As local features the Gabor functions depicted in Fig 5 were used. Only the responses of the first four receptive fields (top row in Fig. 5) are shown.

Author's personal copy

individual image. In the case of speech such a normalization is more difficult and has to be adaptive. For the time being we omitted the binarization but we are convinced that such an adaptive input normalization followed by stronger non-linearities will be very beneficial for the feature extraction process.

Fig. 6 visualizes the results of the first feature extraction layer with the example from Fig. 2. The preference of the receptive fields for certain orientations can clearly be seen. One can further notice that as a consequence of the Winner-Take-Most competition mainly only one feature is active at a given point in the spectrogram. In the following we will refer to these features as $\boldsymbol{c}^{(1)}$.

4.3. Extraction of combination features

In contrast to previous approaches which extract spectro-temporal information we only rely on a few initial neurons in the first layer $Q^{(1)}$ but introduce a second layer $Q^{(2)}$ which combines the responses of the neurons on the first layer to more complex spectro-temporal patterns.

Each of the $n^{(2)}$ combination patterns in $Q^{(2)}$ is composed of $n^{(1)}$ receptive fields $\boldsymbol{w}_{l,k}^{(2)}$, i.e. one for each of the neurons in the previous stage $Q^{(1)}$. The coefficients of these receptive fields are non-negative and span all frequency channels. Similarly to (4) the activity $q_k^{(2)}(t)$ of the kth neuron at the time t is given by

$$q_k^{(2)}(t) = \sum_{l=1}^{n^{(1)}} \left(\boldsymbol{c}_l^{(1)} * \boldsymbol{w}_{l,k}^{(2)} \right)(t,f) \tag{11}$$

with the main difference that the final activity $q_k^{(2)}(t)$ is the result of a summation over all $n^{(1)}$ receptive fields. 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 (11) as both the $\boldsymbol{c}_l^{(1)}$ and the $\boldsymbol{w}_{l,k}^{(2)}$ are non-negative.

The combination patterns were also learned in an unsupervised manner using Non-Negative Sparse Coding (NNSC) (Hoyer, 2004). NNSC differs from Non-negative Matrix Factorization (NMF) by the presence, in the cost function (12), 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 $\Delta = 40$ ms of the first layer activations $\boldsymbol{c}_l^{(1)}$. From these patches we learned $n^{(2)} = 50$ combination features by minimizing the following cost function (Wersing and Körner, 2003):

$$E = \sum_i \left\| \boldsymbol{P}_i - \sum_{k=1}^{n^{(2)}} \alpha_{k,i} \boldsymbol{w}_k^{(2)} \right\|^2 + \beta \sum_i \sum_{k=1}^{n^{(2)}} |\alpha_{k,i}|, \tag{12}$$

where $\boldsymbol{P}_i$ is a tensor representing the $n^{(1)}$ layers of the ith patch, the $\boldsymbol{w}_k^{(2)}$ are $n^{(2)}$ non-negative tensors each of them containing the $n^{(1)}$ receptive fields $\boldsymbol{w}_{l,k}^{(2)}$, the $\alpha_{k,i}$ are nonnegative reconstruction factors, and β is a parameter allowing to control the sparsity of the learned features.

4.4. Integration with an HMM

HMMs are still the dominant paradigm for speech recognition. Therefore, we also use them to assess the performance of the proposed feature extraction framework. Due to notorious shortages in training data it is advisable to use diagonal covariance matrices for the Gaussian mixtures. This entails that decorrelated features are better suited. Consequently, we perform a Principal Component Analysis (PCA) prior to transferring the features to the HMM. The PCA statistics were determined from the training part of the TIDigits database. Additionally, we also add Delta and Delta–Delta features (i.e. the first and second derivative of the features) to the feature vector. The integration of the Delta features is done prior to the calculation of the PCA. When in the following we refer to HIST features we refer to the features as obtained after the application of the PCA.

5. Results

To asses the performance of the proposed HIST features we use two different datasets. In both cases the task is the robust speaker-independent recognition of continuous digits under a comprehensive variety of additional background noise. This setup is inspired by the Aurora-2 framework (Pearce and Hirsch, 2000). The difference between the two datasets is that our first and main dataset consists of wideband, i.e. 16 kHz, signals whereas Aurora-2, our second dataset, consists of telephone quality signals, i.e. 8 kHz bandwidth with additional channel distortions. Via different tests on the wideband dataset we determine the role the different elements of the presented feature extraction framework play for the overall performance. The evaluation on Aurora-2 serves to further substantiate our results.

5.1. Speech databases

5.1.1. Wideband corpus

Similar to Aurora-2 we also derived our wideband corpus from TIDigits. TIDigits contains 326 speakers each pronouncing 77 digit sequences (Leonard et al., 1984). In contrast to Aurora-2 we use the full TIDigits set. To this data we added four types of noise from the Noisex database (Varga and Steeneken, 1993): White, Babble, Factory, and Car. Each noise type was added at Signal to Noise Ratios (SNRs) ranging from −5 dB. . . inf, i.e. we also kept the clean signal. The differences in our speech data to Aurora-2 are:

- We downsampled signals to 16 kHz instead of 8 kHz.
- When mixing the signals with noise using FaNT (Hirsch, 2005) we used the G.712 only for the noise and signal

Author's personal copy

level estimation, i.e. the obtained signals have no channel distortions.
- We took four types of noise from the Noisex database.
- Each individual test condition contains significantly (12 times) more utterances than Aurora-2.

We have chosen a sampling rate of 16 kHz for our main corpus as our primary application domain is the speech based interaction with Honda's humanoid robot. In this case a restriction to 8 kHz is without justification. Additionally, also telephone speech has already partially moved to higher bandwidths.

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

5.1.2. Aurora-2

Aurora-2 uses a subset of 8440 training sentences out of the 12,549 training sentences contained in TIDigits. These signals were downsampled to 8 kHz and convolved with a G.712 characteristic yielding telephone quality signals with an effective bandwidth of 0.3, ..., 3.4 kHz. Aurora consists of a clean and a multi condition training part, both of equal size. In the multi condition training part recordings from inside a subway, babble noise, car noise, and recordings in an exhibition hall are added to the signals. The first test data set contains signals degraded by the same types of noise as used in the multi condition training. In the second set recordings from a restaurant, in a street, at an airport, and at a train station are added to the signals. Finally, in the third set the signals are convolved with a MIRS instead of a G.712 characteristic. As noise types subway and street are used for this set. All noise types were added, as in our wideband dataset, at SNR levels ranging from -5 dB... inf. Each individual test contains only 1001 utterances, compared to 12,547 in our wideband dataset. As a consequence the statistical significance of the results obtained on Aurora-2 is notably inferior to that we obtained on our wideband dataset. If not stated otherwise, all following tests are performed on the wideband dataset.

5.2. Parameterization of the feature extraction framework

The proposed HIST feature extraction framework requires the proper setting of different parameters. With the exception of those experiments where we mention a different setting of the parameters the following parameterization is used for all subsequent tests.

For the extraction of local features on the first level of the hierarchy we used 8 filters learned via ICA. The corresponding receptive fields are depicted in Fig. 4. The sizes of the receptive fields were 20×20 at a sampling rate of 400 Hz. Hence they span 50 ms and 20 channels.

For the competition between the different local features we used $\gamma^{(1)} = 0.7$. The following non-linear compression uses a value of $\vartheta^{(1)} = 0.25$. In the pooling stage the dimensions in time and frequency are reduced by four.

We learned $n^{(2)} = 50$ combination features $(q_1^{(2)}(t), \ldots, q_{n^{(2)}}^{(2)}(t))^T$ all spanning the full frequency range. As temporal extend we used 40 ms wide receptive fields. At a sampling rate of 100 Hz this corresponds to four samples.

Delta (resp. Delta–Delta) features were computed using a 9th order FIR low-pass (resp. band-pass). After decorrelation of the feature vectors via PCA we retained from the 150 features the 39 with the highest eigenvalues. We have chosen this number as it corresponds to the dimensionality of the MFCC features we use as a benchmark.

Due to the several non-linearities in our processing the performance gain due to a modification at a certain stage in the processing is a very unreliable predictor for the performance of the complete hierarchy. For this reason we always report the performance of the whole hierarchy. An exception to this is of course the investigation on the contribution of the different layers.

5.3. Comparison to purely spectral features

First, we want to compare the results we achieve with the proposed HIST features to those of conventional spectral features. As benchmark we used Mel Frequency Cepstral Coefficients (MFCCS) (Davis and Mermelstein, 1980) and RelAtive SpecTral Perceptual Linear Predictive (RASTA-PLP) features (Hermansky and Morgan, 1994). In the case of MFCC features we applied Cepstral Mean Subtraction as it yielded in almost all cases significant performance improvements. For both feature types we also calculated Delta and Delta–Delta features in the same fashion as for the HIST features. When comparing the recognition scores in Fig. 7 one can see that the HIST features show inferior performance than either the MFCC or RASTA-PLP features in clean and low noise conditions (2.5% for HIST vs. 1.5% for RASTA-PLP in clean).[2] From this it is evident that some important information in the speech signal is not well represented by the HIST features. However, with increasing noise level the purely spectral features deteriorate quickly whereas the performance of the HIST features is more robust. One can also observe that the MFCC s yield better results in clean than RASTA-PLP features (1.2% vs. 1.5%) but deteriorate faster when additional background noise is present.

Based on the observation that the performance of the proposed HIST features and the conventional features is complementary, i.e. conventional features showing good performance at low noise levels and HIST features in high noise levels, we combined the HIST features with

[2] We also observed that the HIST features have a strong tendency to produce erroneous word insertions. To counterbalance this we used a word insertion penalty of `-p -70` in the HTK decoding for all results we report for the HIST features alone. All remaining tests, including the combination of HIST with other features, were performed with the standard setting `-p 0`.

Author's personal copy

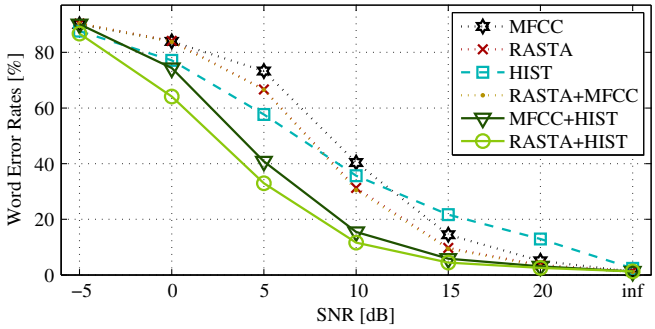


Fig. 7. Word error rates (WERs) when factory noise was added to the speech data.

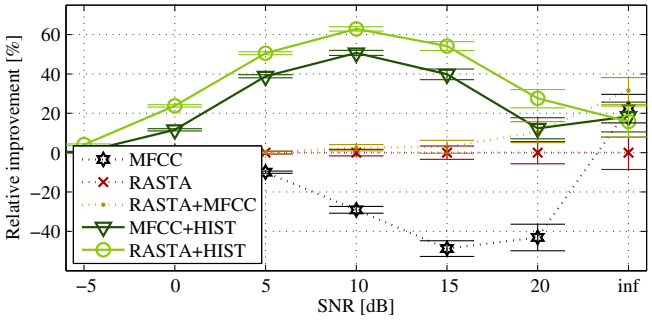


Fig. 8. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

Table 1
Word error rates averaged for an individual type of additional noise over SNR values ranging from −5 dB... inf.

Features	Noise type			
	White	Factory	Babble	Car
MFCC	49.1	44.1	35.6	25.4
RASTA-PLP	43.1	41.0	35.0	19.5
$c^{(1)}$	52.7	51.7	63.0	21.8
HIST	43.3	42.1	53.9	17.9
RASTA-PLP + MFCC	43.5	40.7	32.8	20.4
MFCC + HIST	33.6	33.0	36.4	11.9
RASTA-PLP + $c^{(1)}$	30.6	31.6	37.3	12.0
RASTA-PLP + HIST	26.7	29.1	48.7	10.6
$\text{HIST}_{\text{Gabor}}$	50.4	40.9	57.4	24.1
RASTA-PLP + $\text{HIST}_{\text{Gabor}}$	46.9	39.4	57.4	21.2
$\text{HIST}_{\text{No WTM}}$	58.6	70.3	89.5	47.7
RASTA-PLP + $\text{HIST}_{\text{No WTM}}$	33.8	35.1	45.0	23.7
$\text{HIST}_{20\text{ms}}$	44.2	42.7	51.4	18.9
RASTA-PLP + $\text{HIST}_{20\text{ms}}$	34.9	35.5	40.7	13.9
$\text{HIST}_{80\text{ms}}$	54.0	48.1	54.5	17.1
RASTA-PLP + $\text{HIST}_{80\text{ms}}$	41.7	40.7	60.9	11.7

RASTA-PLP features. Thereby we concatenated at each frame the 45 RASTA-PLP features and the 39 HIST features to an 84 dimensional feature vector. For comparison we also combined MFCC and HIST features and MFCC and RASTA-PLP features in the same way. As can be seen from Fig. 7 the combination of RASTA-PLP and MFCC features gives a significant improvement for clean (1.0%) but does not help when additional background noise is present. In fact, it is hard to distinguish the two curves in Fig. 7. The combination of HIST and RASTA-PLP features shows a very similar performance to the RASTA-PLP features alone in low noise conditions and gives substantial improvements when the noise level increases. The same is true for the combination of HIST and MFCC features although the overall performance is not as good, most likely due to the inferior performance of the MFCC features with additional background noise. Table 1 displays the recognition results for each of the 4 noise types tested.

To better quantify the performance of the combination of the proposed HIST features with RASTA-PLP features compared to conventional features we visualized the relative improvements over RASTA-PLP features in Fig. 8. Improvements of the combination of HIST and RASTA-PLP features are moderate for clean speech (from 1.5% to 1.3% on clean) and reach levels of 40–60% at medium SNR levels. Over most SNR levels we obtain an SNR gain of approximately 5 dB, i.e. we obtain from the combination of HIST and RASTA-PLP features almost the same word error rates as at 5 dB better SNR levels when using only RASTA-PLP features. It can also be seen that the combination of MFCC and RASTA-PLP features does not lead to a significant improvement. The averaged relative improvements are given in Table 2.

Despite the significant improvements we see from the combination of HIST and RASTA-PLP features in most cases the performance with additional babble noise is quite poor. We attribute this to the preprocessing and the strong non-linearities. The preprocessing very effectively enhances the formant structure. This is also the case when only babble noise is present. Due to the following non-linearities this formant structure is then further emphasized such that it does not much differ anymore from speech at a normal level and generates word or phone hypotheses. This can also be seen from a deeper analysis of the errors. With 20 dB babble noise RASTA-PLP features show 3.3% word errors, 0.9% insertions and deletions and additional 2.4% confusions. On the other hand the combination of HIST and RASTA-PLP features yields 14.7% word errors, 12.6% insertions and deletions and only 2.1% confusions. This means that the combination of HIST and RASTA-PLP features does indeed in more cases recognize the correct word but on the other hand produces tremendously more erroneous word insertions.

The largest improvements we see are for scenarios with high significance for real applications. For the rather simple task of continuous digit recognition word error rates should certainly not exceed 10%. When adding the HIST features to the RASTA-PLP features at an SNR level of 10 dB in white noise, factory noise, and car noise we see improvements form 13.5% to 7.0%, from 31.3% to 11.6%, and from 8.0% to 3.7%. Hence more than a reduction by two down to tolerable error rates.

Author's personal copy

Table 2
Relative word error rates averaged for an individual type of additional noise over SNR values ranging from −5 dB... inf. The error rates obtained with RASTA-PLP features were used as a baseline.

Features	Noise type			
	White	Factory	Babble	Car
MFCC	−29.1	−15.4	2.5	−13.1
RASTA-PLP + MFCC	3.4	6.9	17.7	16.4
MFCC + HIST	19.8	24.6	−11.8	38.8
RASTA-PLP + $\boldsymbol{c}^{(1)}$	30.1	26.2	−45.1	31.9
RASTA-PLP + HIST	36.7	34.1	−109.0	41.0
RASTA-PLP + $\text{HIST}_{\text{Gabor}}$	−51.1	10.4	−267.9	11.8
RASTA-PLP + $\text{HIST}_{\text{No WTM}}$	21.4	18.4	−88.9	−12.7
RASTA-PLP + $\text{HIST}_{\text{20ms}}$	22.7	21.4	−53.7	31.4
RASTA-PLP + $\text{HIST}_{\text{80ms}}$	5.4	7.3	−223.0	39.2

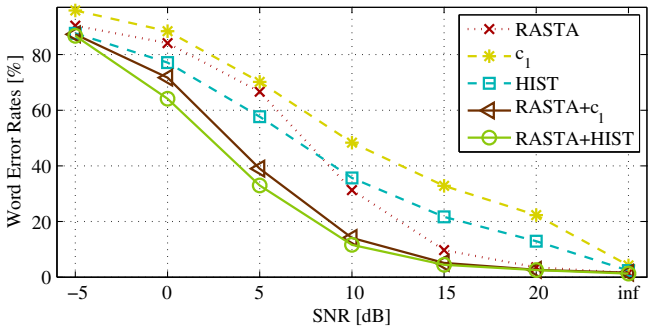


Fig. 9. Word error rates (WERs) when factory noise was added to the speech data.

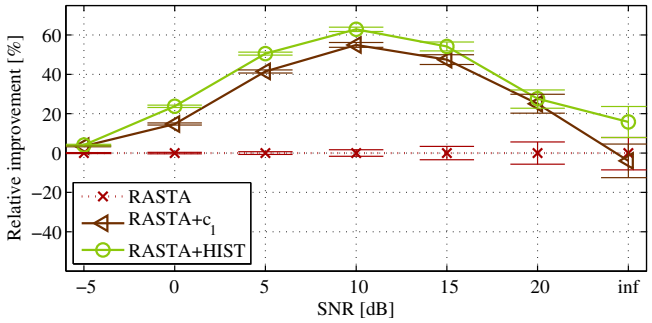


Fig. 10. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

Due to the inferior performance of the MFCC features by themselves and in combination with the HIST features we will in the following only use the RASTA-PLP features.

5.4. Contribution of the hierarchical processing

Next, we want to investigate the contribution of the two hierarchical levels to the overall performance. We calculated for the features $\boldsymbol{c}^{(1)}$ after the first layer of our framework (compare Eq. (10)) a PCA in the same way as for the complete HIST features. With the resulting features and

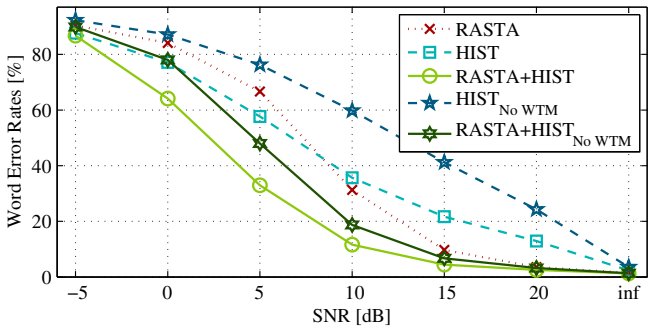


Fig. 11. Word error rates (WERs) when factory noise was added to the speech data.

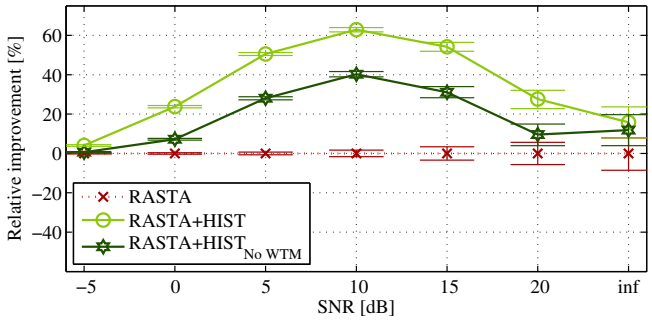


Fig. 12. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

with the concatenation of these features and RASTA-PLP features we trained an HMM. The results indicated with $\boldsymbol{c}^{(1)}$ and RASTA-PLP + $\boldsymbol{c}^{(1)}$ in Figs. 9 and 10 as well as in Tables 1 and 2 show that the features on the first layer alone already yield good performance. Nevertheless the second layer further improves this performance in all cases, except for babble noise. The results based on the $\boldsymbol{c}^{(1)}$ features show lower word insertions than the HIST features when babble noise is added but at the same time yield more confusions (7.1% insertions and deletions and 2.2% confusions compared to 12.6% insertions and deletions and 2.1% confusions at 20 dB SNR).

5.5. Contribution of the feature competition

Now we want to analyze the influence of the competition between the features on the first layer introduced in Eq. (8) on the performance. The results we obtain without the Winner-Take-Most (WTM) competition are displayed in Figs. 11 and 12 as well as Tables 1 and 2. As one can see the impact of the competition depends on the noise type and noise level. In the case of car noise the competition seems to be especially beneficial. However, with additional babble noise the results without the competition are better than with it. A closer look reveals again that without the competition the number of erroneous insertions decreases but at the same time the number of confusions increases. Overall the competition leads to notable improvements.

Author's personal copy

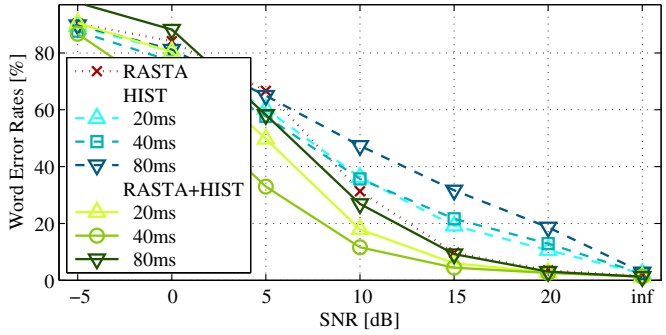


Fig. 13. Word error rates (WERs) when factory noise was added to the speech data.

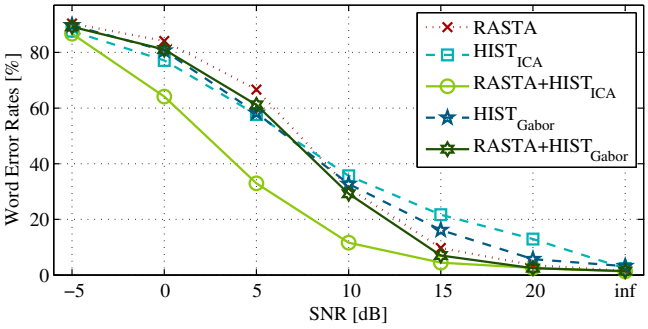


Fig. 15. Word error rates (WERs) when factory noise was added to the speech data.

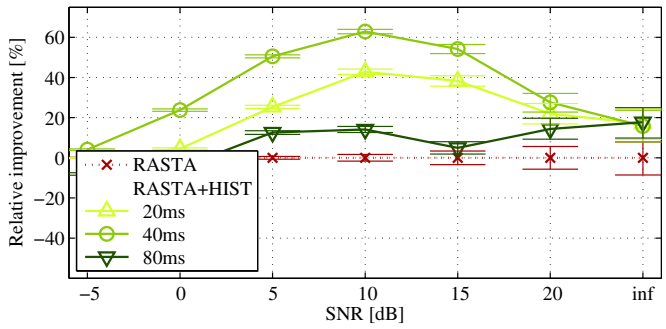


Fig. 14. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

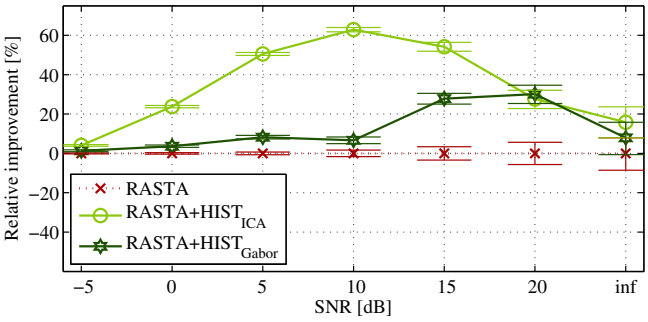


Fig. 16. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

5.6. Contribution of the size of the combination features

One possible source for the improvements we see could be that our features span a longer time window than RASTA-PLP and MFCC features (40 ms compared to 25 ms). To further investigate this we varied the size, i.e. the length, of the combination features. We used 20, 40, and 80 ms. Taking significantly longer time windows would harbour the risk of a strong adaptation of the features to the lexical content. As can be seen from Tables 1 and 2 and Figs. 13 and 14 a width of 40 ms for the second layer seems to be optimal. With smaller and larger widths the performance decreases. Nevertheless with a width of 20 ms the performance is largely better than that of RASTA-PLP features alone. Keeping in mind that already the filters at the first layer have a width of 50 ms we nevertheless conclude from the results that it is not only the longer temporal window which is responsible for the performance improvements we see.

5.7. Choice of the first layer

In this experiment we want to evaluate the role of the receptive field shape on the first layer. In the tests so far we used the features learned via ICA (compare Fig. 4). In Figs. 15 and 16 as well as Tables 1 and 2 we compare them to features derived from the Gabor functions displayed in Fig. 5. Even tough the shape of the Gabor filters seems to match the shapes one observes in the spectrograms much better than the ICA features the results we obtain are significantly inferior to those with the features learned via ICA. During these tests we varied the sizes and frequencies of the Gabor filters to a large extend but did not find a setting which yielded better results as those reported above. This could change when one uses more than just 8 filters.

5.8. Adaptation to HMMs

Standard HMMs yield best results with independent features which statistics can well be modeled as the overlay of a few Gaussian distributions (Hermansky, 1998). However, as can be seen from the feature covariances in Fig. 17(a) and (b) the correlation between the different dimensions of the HIST features is much stronger than for RASTA-PLP features. The PCA step effectively eliminates this problem and hence is indispensable for the successful integration with an HMM featuring diagonal covariance matrices (compare Fig. 17(c)).

In addition, we use the PCA also to reduce the dimensionality of the features. In Figs. 18 and 19 we compare the results we obtain when we either use all 150 dimensions of the HIST features and combine them with RASTA-PLP

Author's personal copy

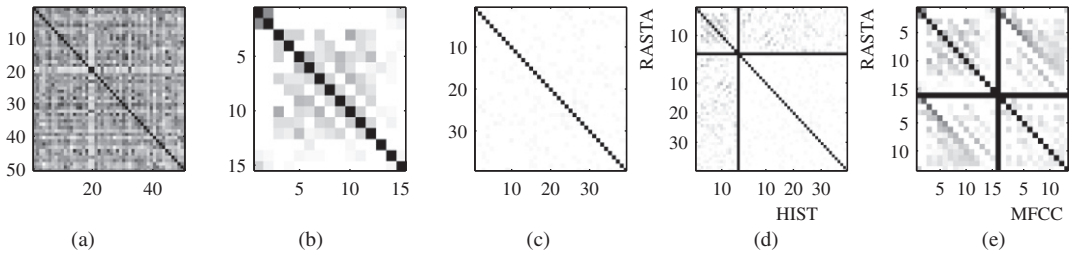


Fig. 17. Covariance matrices of HIST prior to PCA (a), RASTA-PLP (b), HIST after PCA (c), RASTA-PLP + HIST (d), and RASTA-PLP + MFCC (e) features. Prior to the calculation of the covariance matrices we removed the mean and performed a variance normalization of each feature individually. Delta and Delta-Delta features were excluded, except for the HIST features after the PCA in (c) and in combination with RASTA-PLP features in (d). For the combination of RASTA-PLP + HIST and RASTA-PLP + MFCC the separation between the two feature sets is highlighted with a black bar.

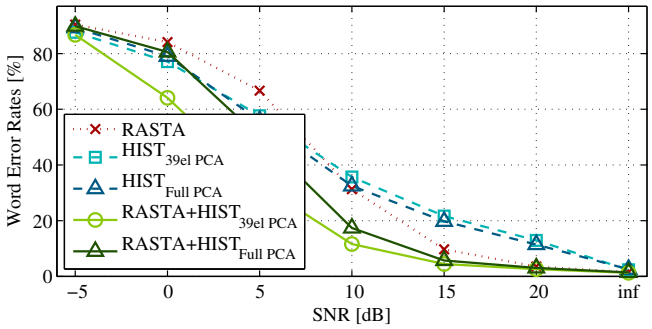


Fig. 18. Word error rates (WERs) when factory noise was added to the speech data.

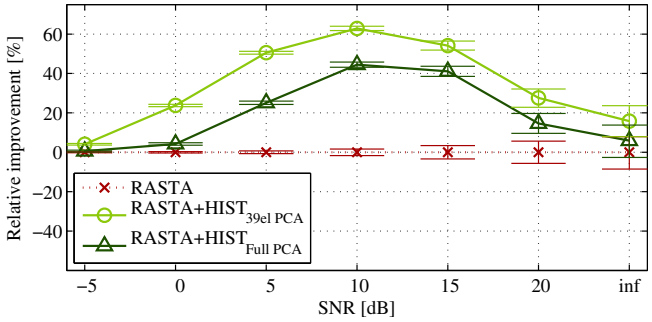


Fig. 19. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

features to those when we reduce the dimensionality to 39 as we did in the previous experiments. As can be seen from these figures the performance slightly improves for the HIST features alone but does decrease for the combination of HIST and RASTA-PLP features. We attribute this to the weakness of the HMMs to identify the relevant features in this combination. When we use all 150 HIST features the additional features resulting from eigenvectors with low eigenvalues and putatively low significance seem to partly mask the 45 RASTA-PLP features. Hence, the preselection of the relevant HIST features by retaining only the 39 features with the highest eigenvalues seems to be beneficial when combining them with lower dimensional features in an HMM recognition system.

5.9. Complementarity of the features

We also want to use the covariance matrices between the different feature types to shed some light on the causes for the improvements we see. Fig. 17(d) and (e) show that the correlation between the HIST features and the RASTA-PLP features is much lower than that between the MFCC features and the RASTA-PLP features. From this it follows that the HIST features deliver information complementary to RASTA-PLP features whereas MFCC and RASTA-PLP features extract very similar information. This explains on one hand why we do not see noticeable improvements when combining RASTA-PLP and MFCC features and on the other hand why we do observe them when combining either RASTA-PLP or MFCC features with HIST features.

Table 3
Word error rates averaged for an individual type of additional noise over SNR values ranging from −5 dB... inf when training was performed on noisy data.

Features	Noise type			
	White	Factory	Babble	Car
RASTA-PLP	27.5	26.4	22.0	6.5
HIST	24.8	27.7	40.0	8.4
$HIST_{adapted}$	26.2	28.7	49.8	16.5
RASTA-PLP + HIST	23.9	24.9	22.9	5.3
RASTA-PLP + $HIST_{adapted}$	21.9	23.8	22.8	6.2

5.10. Matched training

In all previous tests the features as well as the HMM models were trained on clean data and tested with different levels of additional background noise. In most scenarios one has some a priori information on the type and strength of the background noise one will face. Training the HMM models with speech corrupted by this additional background noise significantly increases performance when later tested in similar conditions. Therefore, we additionally performed a test where we added all four noise types at SNR levels of 20 and 10 dB to the training set of

Author's personal copy

Table 4
Relative word error rates averaged for an individual type of additional noise over SNR values ranging from −5 dB... inf when training was performed on noisy data. The error rates obtained with RASTA-PLP features were used as a baseline.

Features	Noise type			
	White	Factory	Babble	Car
RASTA-PLP + HIST	21.7	9.7	2.1	17.6
RASTA-PLP + HIST$_{adapted}$	30.3	19.1	4.1	11.0

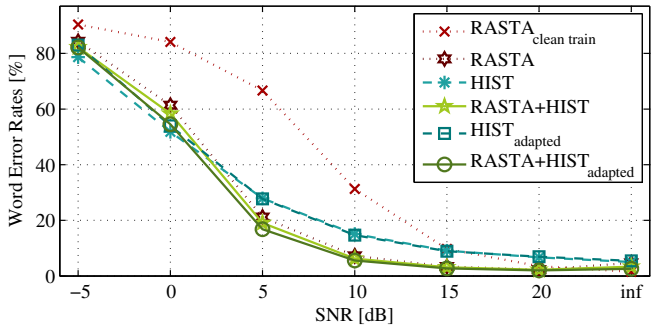


Fig. 20. Word error rates (WERs) when factory noise was added to the speech data. Training was performed on noisy data. Results for the original HIST features or HIST features adapted to the training condition are displayed.

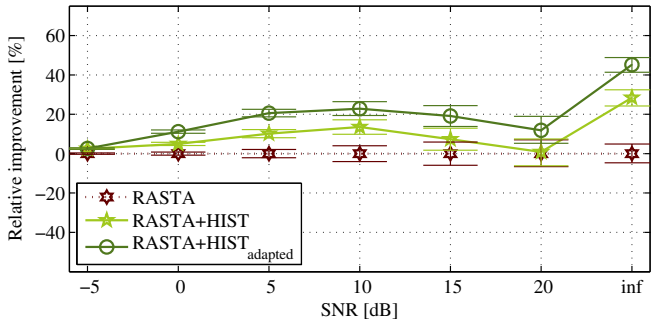


Fig. 21. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Training was performed on noisy data. Results for the original HIST features or HIST features adapted to the training condition are displayed. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

TIDigits. We combined this with the clean signals and trained the HMM models. For computational reasons we used in all cases only a third of the training set. During recognition we used the same data as in the previous tests.

As can be seen from Tables 3 and 4 as well as Figs. 20 and 21 this matched training reduces error rates in well matching scenarios to a large extend. This is true for RASTA-PLP features as well as for the combination of RASTA-PLP and HIST features. On the other hand the performance on the clean signal severely decreased (from 1.5% to 4.7% with RASTA-PLP and 3.4% for the combination of HIST and RASTA-PLP). What we also observed is that the strong tendency of the HIST features to generate insertion errors in the case of babble noise disappeared. Now we see an improvement from the combination of RASTA-PLP and HIST features for all cases tested. We attribute this to the fact that in this case the HMMs learned to reject weak speech hypotheses generated by the HIST features. Overall the improvement from the combination of the features is smaller than in the case when training was performed only on clean signals. The reason for this seems to be two-fold. First, the weaker generalization capabilities of the RASTA-PLP features is remedied by the better adaptation of the HMMs to the scenarios. Second, in this test we used so far features which were learned on clean signals.

In a further test we also learned the features on the noisy data used to train the HMMs. We want to refer to the resulting features as HIST$_{adapted}$. The results in Tables 3 and 4 as well as Figs. 20 and 21 show that this decreased the performance of the HIST features when taken alone but improves performance when used in combination with RASTA-PLP features for almost all cases. We see improvements of around 20%. In the case of clean the improvement even amounts to 45% up to a word error rate of 2.6%. An exception is the case of car noise. We saw before that car noise has a much smaller effect on the recognition performance than the other noise types. Therefore, we presume that including the other noise types in the training leads for tests with additional car noise to a stronger mismatch between training and testing than in the case of training in clean and testing in noise.

5.11. Evaluation on Aurora-2

The tests so far were all performed on the original TIDigits corpus where we added noise at different SNR levels, i.e. our wideband corpus with a 16 kHz sampling rate. Despite the fact that we compared our features to the commonly used RASTA-PLP and MFCC features we wanted to further substantiate our results with additional tests on Aurora-2.

For the tests on Aurora-2 we only changed the parameters in our feature hierarchy directly related to the new sampling rate of 8 kHz. Thereby the different characteristic of the telephone quality signals also in the low frequency range is not taken into account. We adapted the Gammatone filter bank such that the filter center frequencies are distributed in the range of $0,\ldots,4$ kHz instead of $0,\ldots,8$ kHz as before. The DoG filters from Eq. (1) were adapted accordingly. All remaining parameters remained unchanged.

In Table 5 the results of these test are depicted. Additionally, in Fig. 22 the absolute word error rates and in Fig. 23 the relative word error rates compared to the RASTA-PLP features for subway noise in the clean training condition are given. Figs. 24 and 25 show the corresponding results for multi condition training.

The results show that we obtain substantial improvements for some noise types but there are also problematic

Author's personal copy

Table 5
Word error rates (WERs) and relative improvements for the Aurora-2 dataset. The training was carried out either with clean speech, or with speech mixed with various noise types (multi condition training). Tests were always performed using a mixture of noisy and clean signals. The presented scores are averages over all SNRs levels (i.e. −5 dB... inf).

		Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Train-station	Subway (MIRS)	Street (MRIS)	Average
Clean condition training	RASTA-PLP	35.2	33.8	37.4	38.6	34.4	35.5	31.3	35.2	35.2	35.8	35.2
	HIST	52.5	68.3	51.6	57.1	61.2	42.5	52.0	52.9	36.1	36.4	51.1
	RASTA-PLP + HIST	30.0	34.9	34.0	35.8	36.0	30.5	29.0	30.9	30.9	32.9	32.5
	Rel. improvement	22.3	−11.6	7.5	11.0	−2.9	19.6	6.6	14.8	20.5	8.3	9.6
Multi condition training	RASTA-PLP	18.2	20.9	22.7	21.3	20.4	20.8	18.9	22.2	18.4	20.8	20.5
	HIST	25.0	37.5	24.4	27.0	35.5	27.9	28.6	29.2	22.9	27.9	28.6
	RASTA-PLP + HIST	17.1	20.4	21.8	20.4	19.9	19.0	18.2	20.1	17.0	20.0	19.4
	Rel. improvement	13.2	1.1	10.5	14.8	10.2	16.4	3.2	14.3	16.1	11.9	11.2

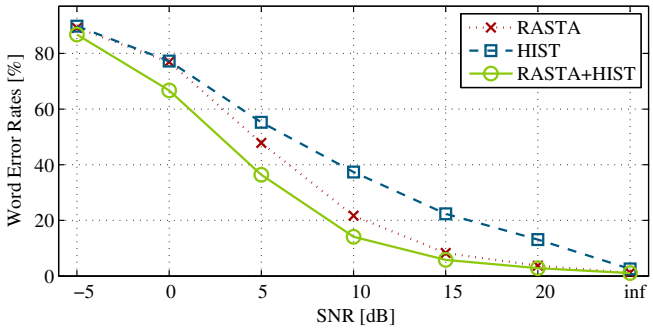


Fig. 22. Word error rates (WERs) obtained on the Aurora-2 dataset with training on the clean training set and testing with additional subway noise.

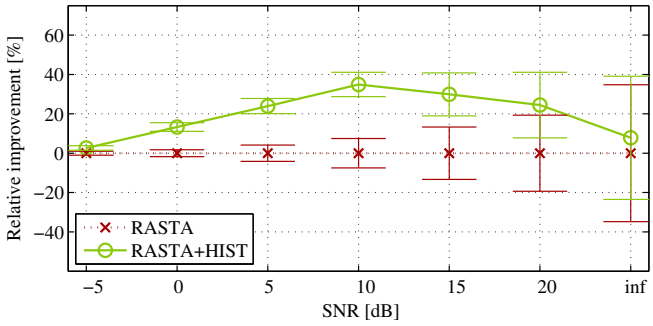


Fig. 23. Word error rates (WERs) obtained on the Aurora-2 dataset with training on the clean training set and testing with additional subway noise relative to those obtained with MFCCS alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

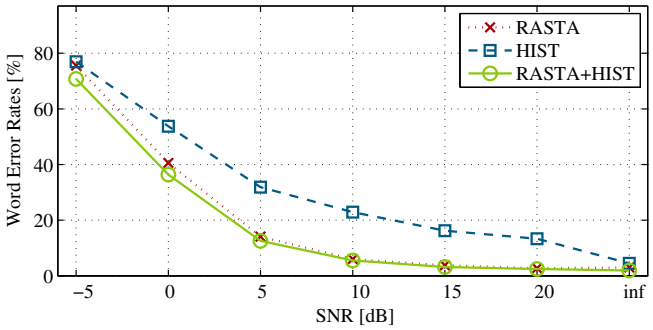


Fig. 24. Word error rates (WERs) obtained on the Aurora-2 dataset with training on the multi condition training set and testing with additional subway noise.

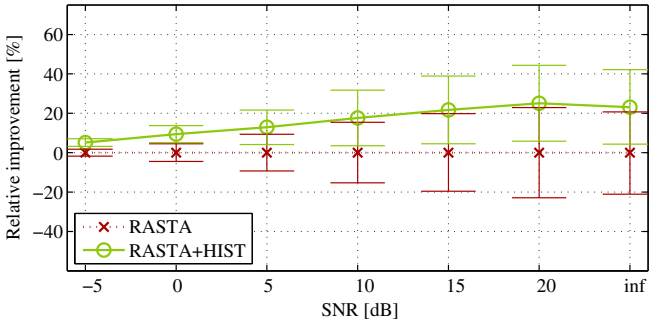


Fig. 25. Word error rates (WERs) obtained on the Aurora-2 dataset with training on the multi condition training set and testing with additional subway noise relative to those obtained with MFCCS alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

noise types where the combination of HIST and RASTA-PLP features yields only small improvements or is even not beneficial. A closer analysis of the noise used in Aurora-2 reveals that additional speech is present in almost all noise types. Only subway, car, and train-station noise do not contain any additional background speech. Hence we assume that the partly unfavorable behavior of our HIST features for most of the remaining noise types is related to the same problem we saw already for babble noise. The results obtained on the multi condition training, depicted in Table 4 as well as Figs. 24 and 25 support this hypothesis. In this case both the HIST features and the HMM models were learned form the multi condition training part of Aurora-2. As we saw already previously, e.g. in Fig. 20, this multi condition training improves performance of the RASTA-PLP features in noisy conditions substantially and at the same time the benefit from the combination with the HIST features is reduced. However, the unfavorable behavior of the HIST features when speech is also present in the noise also gets smaller. In this case

Author's personal copy

we now see again improvements for all types. When comparing these results to the previous ones one hast also to take into account their much weaker statistical significance (this can e.g. be assessed when comparing the confidence intervals depicted in Figs. 8 and 22.

5.12. Computational complexity

The proposed feature extraction framework involves several time consuming 2 D convolutions. Nevertheless, we were able to integrate the feature extraction into an online speech recognition system (Heckmann et al., 2009). We used a Linux workstation with eight Intel Xeon X5355 CPUs working at 2.66 GHz with 4 MB cache and 4 GB RAM per processor. Using only one core of one of the processors, the extraction of the HIST features for 1 s of speech lasts approximatively 280 ms, i.e. 3.5 times faster than real time. Additionally, the higher dimensionality of the resulting feature vector (84 dimensions for HIST + RASTA-PLP compared to only 45 for RASTA-PLP alone) increases the computational load during decoding. However, this can easily be handled by current computers for most speech recognition tasks.

6. Discussion

Similar to other approaches we apply pattern matching techniques inspired by image processing to the task of speech feature extraction. This was motivated not only by the observation that speech shows distinct patterns in the spectrogram but also by recent findings in neurobiology on similarities in the processing in the visual and auditory cortex. We extended this idea by introducing additional concepts from models of visual object perception as unsupervised feature learning, feature competition, hierarchical processing, and high-dimensional sparse representations.

With the results obtained from the first layer of our feature extraction framework we could replicate previous findings by, e.g., Kleinschmidt and Gelbart (2002) as well as Meyer and Kollmeier (2008), that adding spectro-temporal information improves the recognition in noise. We cannot fully rule out the hypothesis that the main reason for the improvements we see lies in the nature of our preprocessing. However, in (Gläser et al., 2010) we compared a preprocessing very similar to the one presented here on a formant tracking task to a Linear Predicitive Coding (LPC) based preprocessing. There we saw that our preprocessing shows under some conditions, especially for the tracking of the second formant in white noise, superior performance but in most cases yields similar results as the LPC based preprocessing. Therefore, we assume that the improvements we see here should be attributed to the representation of spectro-temporal information via the proposed hierarchy. In light of the good results (Kim and Stern, 2010) obtained by applying a 15th-root on the envelope signal we will in the future also investigate if such a strong non-linearity is also beneficial to further improve our preprocessing.

At least for the small number of features we use the actual shape of the features in the first layer seems to play an important role. Despite numerous variations we always achieved significantly better results when learning the features via ICA than when using Gabor shaped features. Hence, the unsupervised data-driven learning of the features seems to be superior to a predefined set of features.

When investigating the feature competition we saw consistent improvements. However, we applied the framework without further modification when calculating the results reported above without the competition. It is well possible that tuning of the parameters will reduce this difference in performance. Nevertheless, we saw in all cases we tested a gain from the competition. On the other hand we also observed that the competition is so far not ideally adapted to the task. In some cases the receptive fields responses largely varied with minor displacements in the spectrogram. As a consequence different features dominate at closely neighboring points thereby breaking up larger structures in the spectrogram and dispersing them over the different features. Modifying the purely local competition such that also surrounding regions are taken into account seems to be a promising approach to further improve the results of this competition step.

One key aspect of our feature extraction framework is the processing in two hierarchical layers. In contrast to purely spectral features and previous spectro-temporal features the proposed second layer, which we termed combination layer, is able to represent complete formant configurations and model non-stationary patterns. We showed that adding this second layer further improves the recognition performance on clean data and in noise. This underlines that the information represented by this second layer is indeed relevant.

In addition to the recognition experiments we demonstrated via a correlation analysis that conventional spectral features as MFCCS and RASTA-PLP features capture similar information. On the other hand, the proposed HIST features extract additional, i.e. complementary, information. From this analysis we also saw that without an additional PCA the different feature dimensions of the HIST features are much stronger correlated than it is the case for RASTA-PLP or MFCC features. Such a behavior is particularly inapt for the use with an HMM. In general one can expect that due to a "coevolution" of spectral features and HMMs an HMM is not very well suited to be used with such novel features (Morgan et al., 2005). Similar to (Meyer and Kollmeier, 2008; Wang et al., 2008; Hermansky and Sharma, 1998; Chen et al., 2004) which use less conventional features and hence features less well suited to HMMs we also expect further improvements when applying alternative recognition backends as e.g. Multilayer Perceptrons or TANDEM modelling (Hermansky et al., 2000; Morgan and Bourlard, 1995). Such alternative recognition backends will be of particular

Author's personal copy

importance when we continue to pursue our goal to obtain sparse and high-dimensional representations and increase the number of elements on the different layers.

We evaluated our features on two different databases each with a comprehensive set of background noise types and SNR levels. On the dataset containing the TIDigits data at a sampling rate of 16 kHz we saw, except for the case of babble noise, consistent improvements from the combination of the HIST features and RASTA-PLP or MFCC features. In particular we obtain strong improvements in scenarios most relevant for actual applications. In the case of babble noise the tendency of the HIST features for high insertions is certainly unfavorable. Also for the tests we performed on the Aurora-2 database we saw unfavorable behavior for those noise types which also included speech. A remedy to this is on one hand the use of additional mechanisms to detect speech on- and offsets and, as we showed, the use of speech corrupted by babble noise already in the training phase. Furthermore, we expect that improvements in our preprocessing will also be able to alleviate this problem.

The improvements we obtained when combining the HIST features with purely spectral features on Aurora-2 were notably lower than those we obtained on TIDigits. One of the reasons for this is certainly that during the development of the HIST features we took only 16 kHz wideband speech data into account and did only change the parameters of the Gammatone filter bank and the DoG filters when applying it to Aurora-2 which uses telephone quality speech sampled at 8 kHz and with an efficient bandwidth of $0.3, \ldots, 3.4$ kHz. Adaptation of the preprocessing, the filter sizes and the different non-linearities will certainly yield to better performance also on telephone quality speech.

When taken alone our features show for good and moderate SNR levels clearly inferior performance to conventional spectral features. From this it follows that some important information captured very well by RASTA-PLP and MFCC features is not represented by the HIST features. We speculate that the smoothing along the time and frequency axis, the preprocessing as well as the pooling stage, could be at the root of this. The information retained might be especially robust against additional distortions but on the other hand not detailed enough to provide precise recognition in undistorted speech. Further experiments will be necessary to validate this hypothesis and develop mechanisms to better capture the relevant information such that the presented approach can serve as a true alternative to conventional features.

So far we only reported results on a continuous digit recognition task. It will be interesting to see how the proposed HIST features perform on more complex tasks and if the features learn only characteristics of the specific task or of speech in general. In a recent experiment we could show that the performance of the HIST features when learned on Timit (Garofolo et al., 1993), a database consisting of phonetically rich American sentences not containing numbers, and then applied on TIDigits the performance only degraded little compared to the case when learning the features on the same dataset as later used for the recognition experiments (Heckmann, 2010). We see this as a first indication that the features are not only learning information limited to the task but more general speech properties.

The additional performance gain we observed via an adaptation of the features to the noise in the matched training condition illustrates that such an adaptation to the current situation can be beneficial. For mammals such task specific plasticity of receptive fields in the auditory cortex seems to play an important role (Fritz et al., 2003). Given that all the learning steps involved in the presented feature extraction framework are unsupervised such an online adaptation seems to the least possible.

Despite the significant improvements we obtained compared to spectral features, many questions on how to optimally combine the different elements we presented here are still open. This is also manifest in the observation we made that by different settings of parameters one could tune the features to have different properties like better performance on clean at the cost of inferior performance in noise, better performance when used alone compared to better performance when combined with RASTA-PLP features and so on. Also for this reason we always reported the performance of the whole hierarchy. Overall, we think that we could show that the different ideas borrowed from neurobiology and vision research we introduced as well as the framework as a whole enable more robust speech recognition. Nevertheless, we see this rather as early steps with many more to follow.

Acknowledgments

First of all we want to thank Heiko Wersing for providing us the algorithms of his object recognition system and many advise on how to use it. Next, we want to thank Stephan Hasler for supporting us with the different learning algorithms. We also want to thank Claudius Gläser for assisting us with the preprocessing and many fruitful discussions. Furthermore, we want to thank Mark Dunn, Bram Bolder, Antonello Ceravola and, Marcus Stein for their help with the computer and software infrastructure. Finally, we want to thank the anonymous reviewers which helped with their constructive comments to improve the paper.

References

Baer, T., Moore, B., Gatehouse, S., 1993. Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times. J. Rehabil. Res. Develop. 30, 49.

Behnke, S., 2003. Discovering hierarchical speech features using convolutional non-negative matrix factorization. In: Proc. Internat. Joint Conf. on Neural Networks (IJCNN), Vol. 4, pp. 2758–2763.

Chen, B., Zhu, Q., Morgan, N., 2004. Learning long-term temporal features in LVCSR using neural networks. In: Proc. 8th Internat. Conf. on Spoken Language (ICSLP). ISCA.

Author's personal copy

Childers, D., Lee, C., 1991. Vocal quality factors: analysis, synthesis, and perception. J. Acoust. Soc. Amer. 90, 2394.

Cho, Y., Choi, S., 2005. Nonnegative features of spectro-temporal sounds for classification. Pattern Recognition Lett. 26 (9), 1327–1336.

Comon, P., 1994. Independent component analysis: a new concept? Signal Process. 36, 287–314.

Crick, F., 1984. Function of the thalamic reticular complex: the searchlight hypothesis. Proc. Natl. Acad. Sci. 81 (14), 4586–4590.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Proc. 28 (4), 357–366.

de Charms, R., Blake, D., Merzenich, M., 1998. Optimizing sound features for cortical neurons. Science 280 (5368), 1439–1443.

Domont, X., Heckmann, M., Joublin, F., Goerick, C., 2008. Hierarchical sectro-temporal features for robust speech recognition. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, Las Vegas, Nevada, pp. 4417–4420.

Domont, X., Heckmann, M., Wersing, H., Joublin, F., Menzel, S., Sendhoff, B., Goerick, C., 2007. Word recognition with a hierarchical neural network. Advances in Nonlinear Speech Processing. Lecture Notes in Computer Science. Springer, Berlin/Heidelberg, pp. 142–151.

Dusan, S., Rabiner, L., 2005. On integrating insights from human speech perception into automatic speech recognition. In: 9th Eur. Conf. on Speech Communication and Technology (EUROSPEECH). ISCA, Lisbon, Portugal.

Elhilali, M., Shamma, S., 2006. A biologically-inspired approach to the cocktail party problem. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, Toulouse, France.

Ezzat, T., Bouvrie, J., Poggio, T., 2007. Spectro-temporal analysis of speech using 2-D Gabor filters. In: Proc. INTERSPEECH. ISCA, Antwerp, Belgium.

Fant, G., 1970. Acoustic Theory of Speech Production. Mouton De Gruyter.

Fant, G., 1979. Glottal source and excitation analysis. Speech Transmiss. Lab. Q. Prog. Stat. Rep. 1, 70–85.

Felleman, D., Van Essen, D., 1991. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1 (1), 1–47.

Fergus, R., Perona, P., Zisserman, A., 2003. Object class recognition by unsupervised scale-invariant learning. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Vol. 2.

Flynn, R., Jones, E., 2008. Combined speech enhancement and auditory modelling for robust distributed speech recognition. Speech Comm. 50 (10), 797–809.

Fritz, J., Shamma, S., Elhilali, M., Klein, D., 2003. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat. Neurosci. 6 (11), 1216–1223.

Fukushima, K., 1980. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernet. 36 (4), 193–202.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., 1993. DARPA TIMIT acoustic–phonetic continuous speech corpus CD-ROM. Philadelphia.

Gläser, C., Heckmann, M., Joublin, F., Goerick, C., 2010. Combining auditory preprocessing and Bayesian estimation for robust formant tracking. IEEE Trans. Audio Speech Lang. Process. 18 (2), 224–236.

Haque, S., Togneri, R., Zaknich, A., 2009. Perceptual features for automatic speech recognition in noisy environments. Speech Comm. 51 (1), 58–75.

Heckmann, M., 2010. Supervised vs. unsupervised learning of spectro temporal speech features. In: Accepted for ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA).

Heckmann, M., Brandl, H., Domont, X., Bolder, B., Joublin, F., Goerick, C., 2009. An audio–visual attention system for online association learning. In: Proc. INTERSPEECH. ISCA, Brighton, UK.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Amer. 87 (4), 1738–1752.

Hermansky, H., 1998. Should recognizers have ears? Speech Comm. 25 (1–3), 3–27.

Hermansky, H., Ellis, D., Sharma, S., 2000. Tandem connectionist feature extraction for conventional HMMsystems. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 3. IEEE, Istanbul, Trukey.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Proc. 2 (4), 578–589.

Hermansky, H., Sharma, S., 1998. TRAPS-classifiers of temporal patterns. In: 5th Internat. Conf. on Spoken Language Processing (ICSLP). ISCA, Sydney, Australia.

Hirsch, G., 2005. FaNT filtering and noise adding tool. Tech. rep., Niederrhein University of Applied Sciences.

Hoyer, P., 2004. Non-negative matrix factorization with sparseness constraints. J. Machine Learn. Res. 5, 1457–1469.

Hubel, D., Wiesel, T., 1965. Receptive fields and functional architecture in two nonstriate visual aeas (18 and 19) of the cat. J. Neurophysiol. 28 (2), 229–289.

Hyvärinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Networks 10, 626–634.

Kim, C., Stern, R., 2010. Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, Dallas, TX, pp. 4574–4577.

King, A., Nelken, I., 2009. Unraveling the principles of auditory cortical processing: can we learn from the visual system? Nat. Neurosci. 12 (6), 698–701.

Klein, D., Konig, P., Kording, K., 2003. Sparse spectrotemporal coding of sounds. EURASIP J. Appl. Signal Process. 2003 (7), 659–667.

Kleinschmidt, M., 2002. Methods for capturing spectro-temporal modulations in automatic speech recognition. Acta Acust. Acust. 88 (3), 416–422.

Kleinschmidt, M., Gelbart, D., 2002. Improving word accuracy with Gabor feature extraction. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP). ISCA, Denver, CO.

Leonard, R., Incorporated, T., Dallas, T., 1984. A database for speaker-independent digit recognition. In: Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 9. IEEE, San Diego, CA.

Lippmann, R., 1997. Speech recognition by machines and humans. Speech Comm. 22 (1), 1–15.

Mesgarani, N., Slaney, M., Shamma, S., 2006. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. IEEE Trans. Audio Speech Lang. Proc. 14 (3), 920–930.

Meyer, B., Kollmeier, B., 2008. Optimization and evaluation of Gabor feature sets for ASR. In: Proc. INTERSPEECH. ISCA, Brisbane, Australia.

Morgan, N., Bourlard, H., 1995. Continuous speech recognition. IEEE Signal Process. Mag. 12 (3), 24–42.

Morgan, N., Zhu, Q., Stolcke, A., Sonmez, K., Sivadas, S., Shinozaki, T., Ostendorf, M., Jain, P., Hermansky, H., Ellis, D., et al., 2005. Pushing the envelope-aside. Signal Process. Mag. IEEE 22 (5), 81–88.

Olshausen, B. et al., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381 (6583), 607–609.

Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., C.Zhang, Allerhand, M.H., 1992. Complex sounds and auditory images. In: Cazals, Y., Demany, L., Horner, K. (Eds.), Auditory Physiology and Perception, Proc. 9th Internat. Symposium on Hearing. Pergamon, Oxford, pp. 429–446.

Pearce, D., Hirsch, H., 2000. The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP). ISCA, Bejing, China.

Rauschecker, J., 1998. Cortical processing of complex sounds. Curr. Opin. Neurobiol. 8 (4), 516–521.

Read, H.L., Winer, J.A., Schreiner, C.E., 2002. Functional architecture of auditory cortex. Curr. Opin. Neurobiol. 12 (4), 433–440.

Riesenhuber, M., Poggio, T., 1999. Hierarchical models of object recognition in cortex. Nat. Neurosci. 2 (11), 1019–1025.

Author's personal copy

Schreiner, C.E., Calhoun, B.M., 1994. Spectral envelope coding in cat primary auditory cortex: properties of ripple transfer functions. Audit. Neurosci. 1 (1), 39–62.

Scott, S.K., Johnsrude, I.S., 2003. The neuroanatomical and functional organization of speech perception. Trends Neurosci. 26 (2), 100–107.

Shamma, S., 2001. On the role of space and time in auditory processing. Trends Cogn. Sci. 5 (8), 340–348.

Sherry, Y., Zhao, N.M., 2008. Multi-stream spectro-temporal features for robust speech recognition. In: Proc. INTERSPEECH. ISCA, Brisbane, Australia.

Slaney, M., 1993. An efficient implementation of the Patterson–Holdsworth auditory filterbank. Tech. rep., Apple Computer Co., technical report #35.

Sroka, J.J., Braida, L.D., 2005. Human and machine consonant recognition. Speech Comm. 45 (4), 401–423.

Stevens, K.N., 2000. Acoustic Phonetics. MIT Press, Cambridge, MA.

Sur, M., Garraghty, P., Roe, A., 1988. Experimentally induced visual projections into auditory thalamus and cortex. Science 242 (4884), 1437–1441.

van Hateren, J., Ruderman, D., 1998. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. Proc. Royal Soc. B: Biological Sci. 265 (1412), 2315–2320.

Varga, A., Steeneken, H., 1993. 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 Comm. 12 (3), 247–251.

Vilar, J., 2008. Efficient computation of confidence intervals for word error rates. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, Las Vegas, NV, pp. 5101–5104.

Wang, H., Gelbart, D., Hirsch, H., Hemmert, W., 2008. The value of auditory offset adaptation and appropriate acoustic modeling. In: Proc. INTERSPEECH. ISCA, Brisbane, Australia.

Wersing, H., Körner, E., 2003. Learning optimized features for hierarchical models of invariant object recognition. Neural Comput. 15 (7), 1559–1588.

Young, E.D., 2008. Neural representation of spectral and temporal information in speech. Philos. Trans. Royal Soc. B: Biological Sci. 363 (1493), 923–945.