Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition

Marc René Schädler,[a] Bernd T. Meyer, and Birger Kollmeier
Medizinische Physik, Carl-von-Ossietzky Universität Oldenburg, D-26111 Oldenburg, Germany

(Received 9 February 2011; revised 27 February 2012; accepted 28 February 2012)

In an attempt to increase the robustness of automatic speech recognition (ASR) systems, a feature extraction scheme is proposed that takes spectro-temporal modulation frequencies (MF) into account. This physiologically inspired approach uses a two-dimensional filter bank based on Gabor filters, which limits the redundant information between feature components, and also results in physically interpretable features. Robustness against extrinsic variation (different types of additive noise) and intrinsic variability (arising from changes in speaking rate, effort, and style) is quantified in a series of recognition experiments. The results are compared to reference ASR systems using Mel-frequency cepstral coefficients (MFCCs), MFCCs with cepstral mean subtraction (CMS) and RASTA-PLP features, respectively. Gabor features are shown to be more robust against extrinsic variation than the baseline systems without CMS, with relative improvements of 28% and 16% for two training conditions (using only clean training samples or a mixture of noisy and clean utterances, respectively). When used in a state-of-the-art system, improvements of 14% are observed when spectro-temporal features are concatenated with MFCCs, indicating the complementarity of those feature types. An analysis of the importance of specific MF shows that temporal MF up to 25 Hz and spectral MF up to 0.25 cycles/channel are beneficial for ASR.
 [http://dx.doi.org/10.1121/1.3699200]

PACS number(s): 43.72.Ne, 43.60.Lq, 43.60.Hj, 43.72.Ar [CYE] Pages: 4134–4151

I. INTRODUCTION

Decades of research in the field of automatic speech recognition (ASR) brought numerous methods to improve the recognition performance by increasing the robustness against variability of speech signals. Several of these methods are inspired by the principles of human speech perception, which is motivated by the fact that the robustness of human recognition performance exceeds by far the robustness of ASR performance even in acoustically optimal conditions (Lippmann, 1997; Cooke and Scharenborg, 2008; Meyer *et al.*, 2011b). The sources of variability in spoken language can be categorized into extrinsic sources (e.g., background noise, the room acoustics, or distortions of the communication channel) and intrinsic sources, which are associated with the speech signal itself (e.g., the talkers' speaking style, gender, age, mood, etc.). Compared to the human auditory system, ASR was found to be far less robust against both types of variability (Lippmann, 1997; Benzeghiba *et al.*, 2007).

In this study, the focus lies on the improvement of feature extraction by using a set of physiologically inspired filters (Gabor filters), which is applied to a spectro-temporal representation of the speech signal. In order to choose a set of filters suitable for ASR tasks, a filter bank is defined and used to extract a wide range of spectro-temporal modulation frequencies (MF) from the signal, while at the same time limiting the redundancy on feature level.

Most state-of-the-art ASR systems perform an analysis of short-time segments of speech and use spectral slices, typically calculated from 25 ms segments of the signal as feature input. The most successful implementations of such spectral processing are Mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) and perceptual linear prediction features (PLPs) (Hermansky, 1990). These features are usually concatenated with their first and second order discrete temporal derivation (delta and double-delta features) to incorporate information about the temporal dynamics of the underlying signal on feature level. The PLP feature extraction was later refined by performing RASTA (RelAtive SpecTrA) processing, which effectively suppresses temporal fluctuations that correspond to background noise or changes of the transmission channel (Hermansky and Morgan, 1994). The idea of using temporal cues was implemented in form of temporal pattern (or TRAPS) features, which were found to increase robustness of ASR systems in noisy environments (Hermansky and Sharma, 1999). These approaches suggest that both spectral and temporal integration of a spectro-temporal representation of the signal may be useful for speech processing, which has therefore motivated studies that incorporate such spectro-temporal processing for ASR.

From a physiological point of view, it seems worthwhile to feed spectro-temporal features to ASR engines, since several studies indicate that a similar processing is performed by the auditory system: These findings indicate that some neurons in the primary auditory cortex of mammals are explicitly tuned to spectro-temporal patterns. For example, Qiu *et al.* (2003) used specific spectro-temporal patterns to identify spectro-temporal receptive fields (STRFs) in the auditory cortex in cats. An STRF is associated with a particular neuron or a group of neurons; it is an estimate for the

[a]Author to whom correspondence should be addressed. Electronic mail: marc.r.schaedler@uni-oldenburg.de.

spectro-temporal representation of the sound stimulus that optimally "drives" the neuron. More recent findings show that spectro-temporal representations of human speech found in the primary auditory cortex of ferrets are well-suited to distinguish phonemes (Mesgarani *et al.*, 2008). The observation that such information is encoded in auditory processing stages serves as motivation for the explicit use of this type of representation in speech pattern recognition.

Different types of spectro-temporal features for ASR have been investigated in the past. Ezzat *et al.* (2007a) and Bouvrie *et al.* (2008) analyzed spectro-temporal patches with a 2D discrete cosine transform. They used this representation as a tool for speech analysis and for the extraction of robust features. Heckmann *et al.* (2008) and Domont *et al.* (2008) employed spectro-temporal patches to derive STRFs from artificial neurons. Another type of spectro-temporal features originates directly from the modeling of the patterns observed in the STRFs in the auditory cortex in cats.

Qiu *et al.* (2003) modeled these patterns with two-dimensional Gabor functions. This motivated Kleinschmidt and Gelbart (2002) to apply Gabor filters to the problem of ASR, with the aim of explicitly incorporating spectro-temporal cues on the feature level. An example of a two-dimensional Gabor filter is shown in Fig. 1. These filters were also shown to be suitable for the analysis of speech properties [e.g., for the distinction of plosives, fricatives and nasals (Ezzat *et al.*, 2007b)]. Mesgarani *et al.* (2006) found that the use of auditory Gabor features improves classification results for speech/nonspeech detection in noisy environments. The extraction of features requires a set of Gabor filters in order to capture information about spectral, temporal and spectro-temporal patterns.

One of the challenges when applying Gabor filters to speech-related tasks is finding a suitable set of filters from the vast number of parameter combinations and which extracts relevant information from the spectro-temporal representation. Standard back-ends such as Hidden Markov Models (HMMs) using Gaussian Mixture Models (GMMs) often require the components of input features to be decorrelated, and computational restrictions make the use of very large vectors (with more than 1000 components) difficult.

In the past, different methods were proposed to cope with this challenge. Kleinschmidt (2003) and Meyer and Kollmeier (2011a) used a stochastic feature selection algorithm [the Feature Finding Neural Network (FFNN) (Gramss, 1991)] that was initialized with a random set of 80 filters. Based on the performance on a simple recognition task (i.e., isolated digits), filters that were found to decrease ASR performance were discarded and replaced with a new random filter, which eventually resulted in a set that was found to increase the noise robustness for the recognition of noisy digit strings. Improvements over the MFCC baseline were obtained by using Gabor features as input to a Tandem system that consists of an artificial neural net (or multi-layer perceptron, MLP). The MLP transformed the Gabor input features into posterior probabilities for phonemes. These posteriors were then decorrelated and used as input to a conventional GMM/HMM classifier.

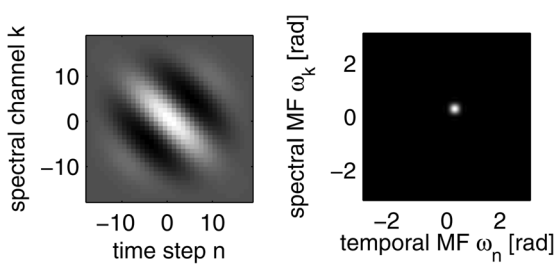

FIG. 1. 2D Gabor filter. (left) Real part. Black and white shading correspond to negative and positive values, respectively. (right) Absolute values of the filter's transfer function in the modulation domain. White shading corresponds to high amplitude.

A different approach is to consider the outputs of the different Gabor filters as feature streams, and start with a very high number of filters (up to tens of thousands compared to the 80 filters mentioned before), and subsequently merging filter outputs that are organized in streams with neural nets. A merger MLP was used to combine isolated streams, and a PCA was applied to its output. This approach was used by Chi *et al.* (2005), Zhao and Morgan (2008), and Mesgarani *et al.* (2010).

These studies have shown that spectro-temporal information helps to increase the robustness of ASR systems. Meyer and Kollmeier (2011a) assumed that the benefits observed for spectro-temporal features (compared to purely spectral feature extraction) arise from a local increase of the SNR since the Gabor functions serve as matched filters for specific spectro-temporal structures in speech, such as formant transitions. However, for several studies (Kleinschmidt and Gelbart, 2002; Meyer and Kollmeier, 2011a), a different database was used for MLP training than for the task for which results were reported, and it is unclear if this additional training material might result in an advantage over setups that do not make use of additional training data. Since all of these studies use the combination of MLPs and PCA, the physical meaning (in terms modulation frequencies) is not directly interpretable from the features that are ultimately fed to the back end. However, when using front-ends as a tool for analysis that might give a hint on what kind of input data is actually helpful, the physical interpretability is a desirable feature.

The aims of this study are to design a filter bank of spectro-temporal filters that are applicable to extract ASR features, and to use these for an analysis of parameters relevant for speech recognition based on spectro-temporal features. Among the design decisions are the number of filters considered for the filter bank, their phase sensitivity, and the spectral and temporal modulation frequencies to be used. Such a 2D filter bank can then be employed to analyze the relative importance of modulation frequencies. Kanedera *et al.* (1999) performed a series of experiments that quantified the importance of purely temporal modulation frequencies for ASR. One of the results is that temporal modulations in the range of 2 Hz to 16 Hz play the dominant role for ASR performance. In this study, this analysis is extended to spectral and spectro-temporal modulation frequencies by performing ASR experiments when specific modulation frequencies are disregarded. Nemala and Elhilali (2010) analyzed the contribution of

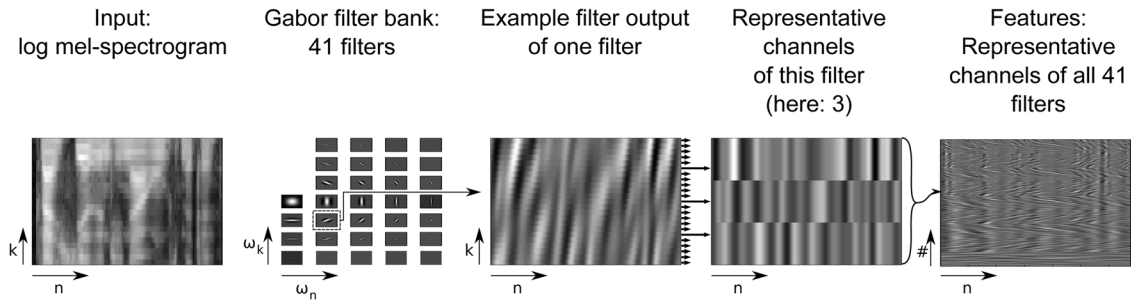

FIG. 2. Illustration of the Gabor filter bank feature extraction. The input log Mel-spectrogram is filtered with each of the 41 filters of the Gabor filter bank. An example filter output is shown. The representative channels of this filter output are selected and concatenated with the representative channels of the other 40 Gabor filters. The resulting 311-dimensional output is used as feature vector.

different temporal and spectral modulation frequencies for robust speech/non-speech classification and found temporal modulations from 12 Hz to 22 Hz and spectral modulations from 1.5 to 4 cycles/octave to be particularly useful to achieve robustness in highly noisy and reverberant environments.

We then evaluate the robustness of these features in the presence of intrinsic and extrinsic sources of variability, and compare them to a range of spectral feature types that are commonly applied in ASR. ASR performance in the presence of additive noise and varying channel characteristics is investigated with two experimental setups (i.e., the widely used Aurora2 digit recognition task that employs the HTK back end, and the Numbers95 task for which a state-of-the-art backend was used). The effect of intrinsic variation is explored using a phoneme detection task (in which phonemes are embedded in short nonsense utterances).

The structure of this paper is reflected by these aims: We first present the design decisions for the Gabor filter bank (Sec. II), how it is applied to feature extraction, and which modulation frequencies were found to be relevant for this ASR task (Sec. II A). Section II C presents the corresponding results. The experiments that investigate the sensitivity of spectro-temporal and baseline features against extrinsic and intrinsic variability are presented in Sec. III B. Sections III C and IV present the results, the discussion and conclusions.

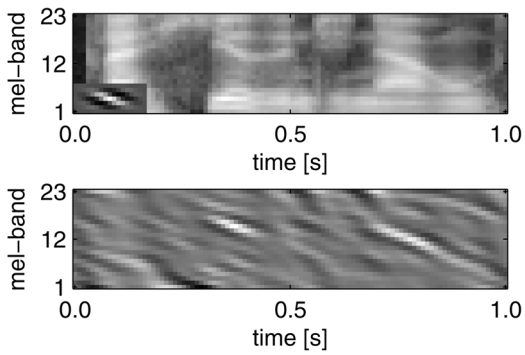

FIG. 3. Illustration of the filtering process with a Gabor filter. (top) Mel-spectrogram of the German sentence "Gleich hier sind die Nahrungsmittel" (The food is right over here) that exhibits spectro-temporal (diagonal) structures that arise from vowel transitions and Gabor filter (real part shown in the lower left corner of the spectrogram). (bottom) 2D filter output obtained by calculating the convolution of the Mel-spectrogram and the real part of the filter. White shading corresponds to high energy on the logarithmically scaled color encoding.

II. GABOR FILTER BANK FEATURES

This section describes the design of the Gabor filter bank, the choice of its parameters, and the calculation of the Gabor filter bank features (GBFB). With these features, we perform an analysis of the importance of phase information in spectro-temporal pre-processing, evaluate the effect of selecting specific modulation frequencies.

A. Calculation of the GBFB features

An overview of the feature extraction scheme with the Gabor filter bank process is illustrated in Fig. 2. First, a Mel-spectrogram is calculated from the speech signal using an implementation of the ETSI Distributed Speech Recognition Standard (ETSI Standard 201 108 v1.1.3 2003). This standard defines the calculation of a Mel-spectrogram that consists of 23 frequency channels with center frequencies in the range from 124 Hz to 3657 Hz. The calculation is based on frames of 25 ms length, while the temporal resolution is 100 frames/s. The spectrogram incorporates a Mel-

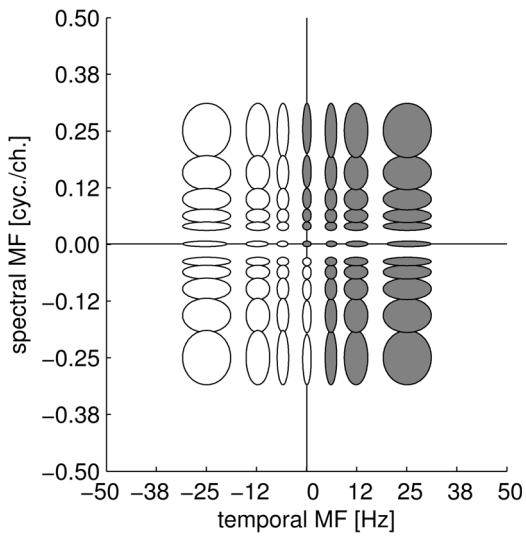

FIG. 4. Illustration of the distribution and size of the transfer functions of the Gabor filter bank filters. Each circle/ellipse corresponds to one Gabor filter and is centered on its center frequency. The circles/ellipses mark the −1 dB level of the filters. With the exception of filters on the axis, the relation between the center modulation frequency and the bandwidth of its passband is proportional. Since only the real part of the filter output is considered for feature extraction, centrally symmetric filters yield identical outputs. Therefore, only the filters that correspond to the filled circles/ellipses are used for feature extraction.

frequency scale that is logarithmic for frequencies above 1 kHz and therefore mimics the mapping of frequencies to specific regions of the basilar membrane in the inner ear. Since the frequency mapping is not strictly logarithmic (with approximately linear at frequencies below 800 Hz), the spectral modulation frequencies are specified in cycles per channel. The absolute output values of the spectrogram are compressed with the logarithm, roughly resembling the amplitude compression performed by the auditory system. The spectrogram is then processed with the filters from the GBFB, which are introduced in Sec. II A 1, by calculating the two-dimensional convolution of the spectrogram and the filter. This results in a time-frequency representation that contains patterns matching the modulation frequencies associated with a specific filter. The filtering process is illustrated in Fig. 3, which shows the original spectrogram, a sample filter, and the filter output.

1. Gabor filter bank

The localized complex Gabor filters are defined in Eq. (1), with the channel and time-frame variables k and n; k_0 denoting the central frequency channel; n_0 the central time frame; ω_k the spectral modulation frequency; ω_n the temporal modulation frequency; ν_k and ν_n the number of semi-cycles under the envelope in spectral and temporal dimension; and ϕ an additional global phase. A Gabor filter is defined as the product of a complex sinusoid carrier [Eq. (1b)] with the corresponding modulation frequencies ω_k and ω_n, and an envelope function [Eq. (1a)].

$$h_b(x) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi x}{b}\right) & -\frac{b}{2} < x < \frac{b}{2}, \\ 0 & \text{else} \end{cases} \tag{1a}$$

$$s_\omega(x) = \exp(i\omega x), \tag{1b}$$

$$g(k_0, n_0, \omega_k, \omega_n, k, n, \nu_k, \nu_n, \phi) = \underbrace{s_{\omega k}(k - k_0)s_{\omega_n}(n - n_0)}_{\text{carrier function}} \cdot \underbrace{h_{\frac{\nu_k}{2\omega_k}}(k - k_0)h_{\frac{\nu_n}{2\omega_n}}(n - n_0)}_{\text{envelope function}} \cdot \underbrace{e^{i\phi}}_{\text{phase factor}}. \tag{1c}$$

For purely temporal and purely spectral modulation filters ($\omega_n = 0$ or $\omega_k = 0$) this definition results in filter functions with infinite support. For that reason the filter size of all filters is limited to 69 channels and 40 time frames. These limits correspond roughly to the maximum size of the spectro-temporal filters in the respective dimensions. Due to the linear relation between the modulation frequency and the extension of the envelope, all filters with identical values for ν_k and ν_n are constant-Q filters.

Since relative energy fluctuations are of special interest for the classification of speech, the DC bias of each filter is removed. This is achieved by subtracting a normalized version of the filter's envelope function from the filter function, so that their DC values cancel each other out. Filters that are centered near the edges of the spectrogram usually do not lie completely within the boundaries of the spectrogram. Hence, the DC removal is applied for all center frequencies separately to avoid artifacts. The effect of the DC removal is that the resulting representation is independent of the global signal energy. Since a removal of the mean on a logarithmic energy scale is the same as dividing by it on a linear scale, this corresponds to a normalization. While cepstral coefficients normalize spectrally, and RASTA processing and discrete derivatives normalize temporally, DC-free Gabor filters naturally normalize in both directions.

The filter bank is designed with the aim of evenly covering the modulation frequencies in the modulation transfer space as schematically illustrated in Fig. 4. Cross-sections of the filter transfer functions along the x axis and y axis of this representation are depicted in Fig. 5.

The distribution of spectro-temporal modulation frequencies is defined by Eq. (2), which ensures that adjacent filters exhibit a constant overlap in the modulation transfer domain. The advantage of this definition is that each filter accounts for a different combination of spectral and temporal modulation frequencies (ω_n, ω_k) and thus has limited correlation with the other filters.

$$\omega_x^{i+1} = \omega_x^i \frac{1 + \frac{c}{2}}{1 - \frac{c}{2}}, \tag{2a}$$

$$c = d_x \frac{8}{\nu_x}. \tag{2b}$$

Figure 5 also explains the meaning of the parameters of the GBFB. The upper and lower bounds for the modulation frequencies are given by $\omega^{\max}$ and $\omega^{\min}$. The width of a filter ω is proportional to the center modulation frequency ω and anti-proportional to ν, which results in constant-Q filters. The distance to the point where two adjacent filters have equal gains (marked with an x) is proportional to the width and the distance factor d. This factor is used to adjust the overlap of adjacent filters, with small values for d resulting in a large overlap and with $d = 1$ corresponding to a coincidence of the first zeros of adjacent filters. The redundancy of the filter outputs due to their overlap can thus be controlled by the distance parameter d.

The modulation frequencies (ω_n, ω_k) can assume positive or negative values. The signs determine the spectro-temporal direction the filter is tuned to. Filters with only one negative modulation frequency correspond to rising spectro-temporal patterns, while other filters correspond to falling spectro-temporal patterns. Since the feature extraction uses the real part of the filter outputs, only filters with positive modulation frequencies and their symmetric versions with

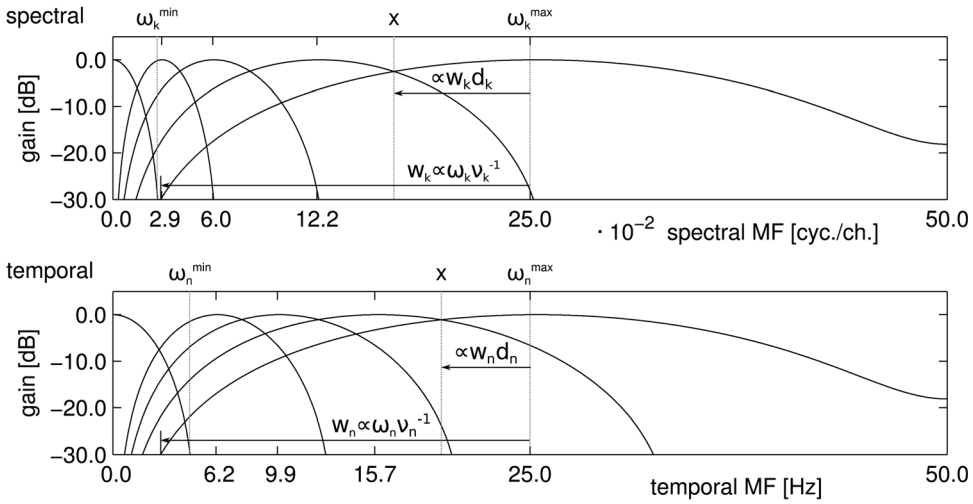

FIG. 5. Cross-section along the spectral and the temporal axis of the modulation transfer space showing the gains of the individual transfer functions. The width of a filter w is proportional to the center modulation frequency ω and anti-proportional to the number of half-waves under the envelope ν, and is indicated here for the highest modulation frequency. The distance to the point where two adjacent filters have equal gains (marked for the filter with the highest modulation frequency with an x) is proportional to the width and the distance factor d. Note that the distance parameter d also controls the overlap between adjacent filters. In the upper panel d_k is chosen 0.3, where in the lower panel d_n is 0.2.

one sign inverted are considered, as inverting both signs would yield identical filters. This relation is illustrated in Fig. 4. Only the filters that correspond to the filled circles/ellises are used. The corresponding filters are depicted in Fig. 6.

2. Selection of representative frequency channels

When using the filter output of all 41 filters, the resulting feature vector is relatively high-dimensional with 23 (frequency channels) × 41 (filters). We reduce the number of feature components by exploiting the fact that the filter output between adjacent channels is highly correlated when the filter has a large spectral extent (cf. Figure 2). Since highly correlated feature components can result in reduced ASR performance (especially when only a small amount of training data is available), a number of representative channels is selected by subsampling the 23-dimensional filter output for each filter. The central channel, corresponding to about 1 kHz, is selected for all feature vectors because the most important cues for ASR are more likely to be found in the center rather than at the edges of the spectrum. Additionally, channels with an approximate distance of a multiple of 1/4 of the filter width to the center channel are included. The value 1/4 is motivated by the sampling theorem in the same way as the minimum window overlap that is needed in a spectrogram for perfect reconstruction.

For filters with the lowest spectral extent, all 23 components are selected for the feature vector, while for the largest filters only a single component (the central frequency channel) is kept. An example with three selected channels is shown in Fig. 2. This selection scheme reduces the filter bank output to 311 dimensions, which is referred to as GBFB features. Alternatively, a principal component analysis (PCA) may be applied to the full filter bank output, which has the same effect as the channel selection (i.e., the decorrelation of feature components, and the reduction of dimen-

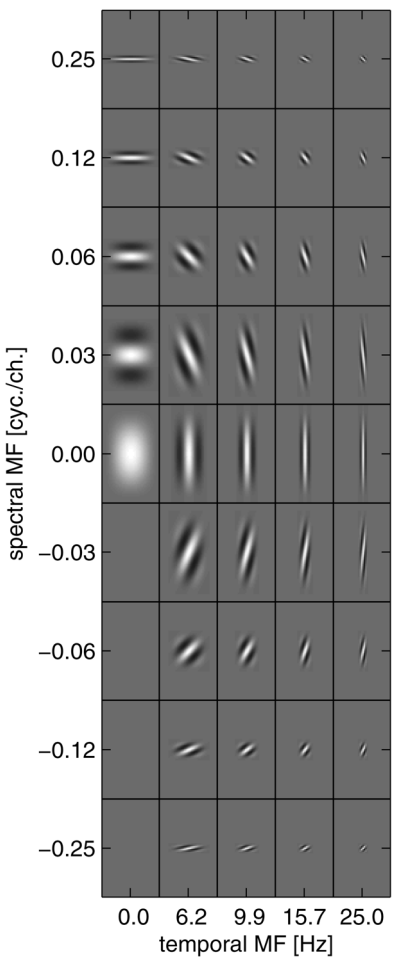

FIG. 6. Real part of the 41 Gabor filters used for the Gabor filter bank feature extraction in time-frequency domain. Black and white shading corresponds to negative and positive values, respectively.

sionality). We therefore test the application of PCA to the filter bank output and compare the results to the proposed scheme of channel selection.

3. Implementation

The calculation of GBFB features results in higher computational load compared to standard front-ends (by a factor of 80 compared to MFCC features), which may be an issue on small-footprint systems. However, GBFB feature calculation can be performed in real-time on a single-core standard PC, and with the current development of dual- and many-core processors, considerable speedups can be achieved by parallelizing the 2D convolutions of the filters. A reference implementation of the GBFB feature extraction in MATLAB is available online (Schädler, 2011).

B. Experiments

Before describing the experiments with the Gabor filter bank, the Aurora 2 framework, the automatic speech recognition framework that is used in all of the following experiments in this section to determine performance and robustness, is introduced.

1. Aurora 2: Digits in noise recognition task

To evaluate robustness against extrinsic variability the Aurora 2 framework is used (Pearce and Hirsch, 2000). It consists of the Aurora 2 speech database, a reference feature extraction algorithm (MFCC), a recognizer setup [Hidden Markov Toolkit (HTK) (Young *et al.*, 2001)], and rules for training and testing. The recognition task is the classification of connected digit strings with artificially added noise. The database contains digits spoken by native English speakers and everyday noise signals recorded at 8 different sites (subway, babble, car, exhibition, restaurant, street, airport, train station). The test set consists of digits with noises added at different SNRs ranging from 20 dB to −5 dB. The standard features used in the Aurora 2 framework are MFCC features with their first and second discrete derivative. For speech-data modeling the HTK recognizer employs Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs).

In the Aurora 2 framework, two training and three test conditions are defined: Clean training uses only clean utterances, while for multi-condition training a mixture of noisy (subway, babble, car, exhibition) and clean digit strings is used. Test set A contains noises also used for training, while for test set B unknown noise types (restaurant, street, airport, station) are used. Test set C contains samples that have been filtered with a different transfer function than the samples of test set A, B, and the training data to simulate a change in communication channel properties.

The HMM back end is configured according to the Aurora 2 guidelines for all feature types: The number of HMM states per word is 18, the number of Gaussian mixtures per state is three; an additional tuning of the back end is not performed. Although tuning might improve results especially when the feature dimension strongly differs from the dimensionality of baseline features, we keep the parameters for reasons of comparability with other studies that use the Aurora 2 framework. For all features the number of time frames is kept constant, because skipping a few frames at the beginning and the end of the utterances improves the performance as it narrows the region to where speech occurs.

The experiments are carried out with different feature types to compare their robustness with respect to the effect of the mismatches between the training and the test data, represented by test sets A, B, and C. The results obtained with the Aurora 2 setup consist of the results for multi-condition and clean training. Word recognition accuracies (WRA) in percent are calculated for each noise condition and for each signal-to-noise ratio (SNR) separately. We also present the relative reduction of the word error rate (WER), which is calculated by determining the relative reduction of error $WER = 1 - WRA$ for each SNR/noise condition (with SNRs ranging from 0 to 20 dB) and averaging over those improvements. The average relative improvements of each noise condition and of test set A, B, and C are calculated to differentiate the effect of different types of mismatches between training and test data. Furthermore, the average word recognition accuracy and the relative improvement for each SNR is calculated.

2. GBFB parameters

This section describes how several of the parameters of the filter bank were chosen. We also compare this choice to the corresponding parameters of the baseline features. Given the structure of the filter bank (that defines, for example, the position of filters in temporal and spectral dimension given a spacing between those filters), we are left with eight parameters that need to be specified: The lowest and highest temporal and spectral modulation frequencies $(\omega_n^{\max}, \omega_k^{\max}, \omega_n^{\min}, \omega_k^{\min})$, the number of periods used for the filters (ν_n, ν_k), and the overlap of adjacent filters (d_k, d_n). The initial values for these parameters are chosen based on the corresponding values of the baseline features (cf. Table I). For instance, the spectral modulation frequencies associated with the baseline MFCCs

TABLE I. GBFB parameter values used for feature extraction in comparison with values derived from parameters of the baseline features.

	Parameters (or their approximated analogues)							
Features	$\omega_k^{\min}\left[\frac{\text{cyc}}{\text{ch}}\right]$	$\omega_k^{\max}\left[\frac{\text{cyc}}{\text{ch}}\right]$	$\omega_n^{\min}$[Hz]	$\omega_n^{\max}$[Hz]	ν_k	ν_n	d_k	d_n
GBFB	0.0254	0.2500	4.38	25	3.5	3.5	0.3	0.2
MFCC WI007	≈0.022	≈0.28	0.0	50	1–13	≈1–3	≈0.13	–
RASTA-PLP	–	–	≈2.6	≈20	–	–	–	–

range from 0.022 to 0.28 cycles/channel, and the parameters for GBFB features were chosen accordingly. The same is true for the temporal modulation frequencies that are relevant in RASTA-processing of signals. Further optimization was carried out by performing a series of ASR experiments varying the parameters one after another on the Aurora 2 task, finding the parameters that result in best overall performance. The optimization was carried out with a fixed phase setting of $\phi = 0$, because this way the maximum amplitude of the filters coincides with the center of the filter independently of its modulation frequency. The parameters were not optimized in any particular order which could have led to finding a local optimum in the parameter space.

To test if GBFB features are overfitted to the Aurora 2 task by the selection of a specific set of parameters, variations of all parameters to the best performing set are evaluated. From the default set of parameters in Table I, each parameter is set to different values, covering a wide range of the plausible parameter space. The selection of frequency channels (Sec. II A 2) was not optimized. Instead we apply the outlined scheme to the full output of the filter bank, and compare the results to transforming the full output with a PCA. The results are presented and discussed in Sec. II C 1.

3. Importance of GBFB phase information

From the output of the Gabor filter bank, either the real or imaginary part, or the absolute values may be used. Using the imaginary part of the output is equivalent to choosing the parameter $\phi = \pi/2$, and effectively using the filters as edge detectors of spectro-temporal events. The absolute values of the output are less sensitive to the exact spectro-temporal location. The phase of the Gabor filters does not matter in this case. To test the importance of the phase information of the filter bank output for robust ASR the performance of the real part, the imaginary part and absolute values of the filter output is compared on the Aurora 2 task. The results are presented and discussed in Sec. II C 2.

4. Relative importance of specific modulation frequencies

In order to evaluate the importance of specific modulation frequencies for ASR, a band-stop experiment is performed that quantifies the contribution of specific combinations of spectral and temporal modulation frequencies (ω_k, ω_n) to the overall ASR performance. For this evaluation, the feature components associated with a specific modulation frequency are removed from the output of the Gabor filter bank. This approach results in 41 different reduced filter sets. Since the number of center frequencies associated with a specific spectro-temporal modulation frequency varies (cf. Sec. II A 2), the number of dimensions removed from the GBFB output ranges from 0 to 22. When the accuracy decreases when omitting filters with a particular modulation frequency, these filters are likely to extract relevant information that is not covered by the remaining Gabor filters. On the other hand, if the accuracy increases when filters are omitted, this indicates that the filters capture information that is either covered by the remaining filters or not relevant for this specific speech recognition task.

The importance of the filters is evaluated with the Aurora 2 task, since this speech material is expected to exhibit a more natural distribution of temporal modulation frequencies compared to the very short utterances from the OLLO database. The Aurora 2 recognizer is trained and tested with each reduced feature representation. The results are presented and discussed in Sec. II C 3.

C. Results and discussion

1. GBFB parameters

Overall recognition performance in % WRA and % relative reduction of the WER over the MFCC WI007 baseline for variations of the GBFB parameters of Table I are presented in Table II. The recognition performance for the GBFB features with altered parameters changes compared to

TABLE II. Overall word recognition accuracies (WRA) and relative reduction of word error rates (relative improvement) compared to the MFCC baseline with clean (c) and multi (m) condition training on the Aurora 2 task for various modifications to the GBFB parameters.

Parameter		ν_k				ν_n				d_k			d_n		
Values		2.5	3.0	4.0	4.5	2.5	3.0	4.0	4.5	0.1	0.2	0.4	0.1	0.3	0.4
WRA [%]	c	56.7	63.0	68.2	70.0	69.6	68.6	62.6	61.2	63.6	65.1	67.4	65.3	67.0	64.8
	m	86.4	88.1	86.5	87.3	83.3	86.7	87.6	87.0	87.1	87.8	87.9	87.8	87.5	84.6
Rel. Imp. [%]	c	−18.8	14.5	33.9	37.7	30.8	33.0	14.3	10.0	22.7	27.0	28.9	28.1	29.4	11.4
	m	−2.8	18.6	9.2	5.27	−48.0	−2.2	13.4	10.2	12.0	14.0	12.5	17.5	11.4	−17.7

Parameter		$\omega_k^{\max} 10^{-2} \left[\frac{\text{cyc}}{\text{ch}}\right]$		$\omega_k^{\max}$ [Hz]		$\omega_k^{\min} 10^{-2} \left[\frac{\text{cyc}}{\text{ch}}\right]$			$\omega_k^{\min}$ [Hz]		
Values		18.75	12.5	18.75	12.5	1.9	3.8	7.61	2.19	3.5	8.75
WRA [%]	c	62.1	61.9	69.3	69.0	66.2	61.4	59.0	67.8	65.5	65.7
	m	86.8	85.0	87.5	88.2	88.1	86.5	89.3	88.9	88.7	87.6
Rel.Imp. [%]	c	16.0	14.2	34.2	33.4	28.4	12.0	−8.3	35.0	28.2	26.3
	m	3.9	−5.0	7.9	7.4	16.6	6.1	24.1	10.5	17.2	15.5

the original set. For some parameters the best values are different for clean and multi-condition training. Hence, the set of parameters that is used for feature extraction is a trade-off between performance for clean training and performance with multicondition training and each could be improved further by selecting different parameters. With many of the changes to the original parameter set, the GBFB features still improve the MFCC WI007 baseline. Some parameters affect more the overall performance, some affect more the relative improvement over the baseline, but there is no clear trend.

The presented results were obtained by selecting frequency channels from the filter output as described in Sec. II A 2. In order to evaluate if a decorrelation and dimension reduction with a PCA should be preferred over channel selection, we apply a PCA either to the full filter bank output (i.e., channel selection is not performed) or the 311-dimensional GBFB features. In each case, the transformation statistics is obtained from the corresponding (clean or multi-condition) training material, and the feature dimension is reduced to 39 (the dimension of the baseline features). From each dimension of the data the mean is removed and the variance is normalized before calculating the PCA coefficients. The results are shown in Table III.

The application of a PCA to the *full* filter bank output results in recognition rates below the GBFB features and the MFCC baseline. When a PCA is applied to the GBFB features with representative channels, the absolute score for clean training is improved, whereas multi-condition results are better with GBFB features. The relative improvements over the baseline are slightly higher with the original GBFB approach. We therefore argue that the direct use of GBFB features should be preferred over PCA-transformed features, since GBFB features are easier to calculate, produce slightly better results on average, and the physical meaning of feature components is retained (i.e., each feature component is associated with a modulation frequency, which enables experiments such as the evaluation of the contribution of such physical parameters to ASR).

The results of the parameter variation and the PCA show that the feature extraction can be optimized for a specific condition. For multi-condition training for example, even less robust patterns may serve for the recognition, as their uncertainty is known. These patterns could be matched by Gabor filters with diverse shapes. In this sense, the GBFB structure limits the fitting to a specific task by greatly reducing the degree of freedom of the feature extraction in contrast to a set of independent Gabor filters. The GBFB features project the log Mel-spectrogram to a over-complete basis of a subspace of the log Mel-spectrogram. The subspace is limited by the lower and upper bounds for the modulation frequencies ($\omega_k^{\min}$ to $\omega_k^{\max}$ and $\omega_n^{\min}$ to $\omega_n^{\max}$). Its degree of over-completeness is adjusted by the distance parameter d and the shape of the basis functions is determined by ν.

It is likely that for different tasks different sets of parameters are optimal, as it is also the case with traditional features. However, we found that none of the parameters of this very generic projection is critical to outperform the MFCC baseline. Nonetheless, ASR systems are non-linear and complex so that the front end and the back end cannot be judged independently. Back ends make strong assumptions about the feature's statistical characteristics, which lead to degraded recognition performance if ignored. A remaining question is if the improvements made with GBFB features for the Aurora 2 task will translate to other ASR setups. For that reason the GBFB features that are adapted to work well with the GMM/HTK back end of the Aurora 2 task are evaluated on another recognition task with a different back end in Sec. III.

TABLE III. Comparison of GBFB features that incorporate the selection of frequency channels from the filter output (GBFB), processing the full filter bank output with a PCA ($^{\text{GBFB}}_{\text{full}}$ and PCA) and application of a PCA to the GBFB features (GBFB and PCA). The recognition performance is presented in word recognition accuracies in % and as relative improvement over the MFCC baseline for clean (c) and multi (m) condition training.

Method PCA		GBFB	GBFB and PCA		$^{\text{GBFB}}_{\text{full}}$ and PCA	
Traindata		–	clean	multi	clean	multi
WRA [%]	c	66.2	**69.6**	64.5	56.5	45.4
	m	**88.1**	82.9	84.6	85.0	86.2
Rel. Imp. [%]	c	**28.4**	28.3	−23.6	−28.6	−79.0
	m	**16.2**	−47.3	14.2	−14.1	−9.2

2. Importance of GBFB phase information

Overall recognition performance in % WRA and % relative reduction of the WER over the MFCC WI007 baseline for the real part, the imaginary part and the absolute values of the GBFB features are presented in Table IV. The accuracies obtained with the real and imaginary part are in the same range, whereas the performance with absolute values (for which the location of spectro-temporal events is smeared out) is reduced considerably. This indicates that phase information is an important factor for ASR, and should be considered in spectro-temporal feature extraction. Since the real-valued filter output performs slightly better than features based on imaginary filters on average, we use the real output for ASR experiments.

3. Relative importance of specific modulation frequencies

In this section, the digit recognition performance is determined based on reduced filter sets, for which a spectro-temporal modulation frequency is omitted as described in

TABLE IV. Average word recognition accuracies (WRA) and relative reduction of word error rates (relative improvement) compared to the MFCC baseline with clean (c) and multi (m) condition training on the Aurora 2 task for the real part, the imaginary part and the absolute values of the GBFB features.

Modification		None (real)	Imaginary	Absolute
WRA [%]	c	66.2	**67.0**	48.8
	m	**88.1**	87.8	81.8
Rel. Imp. [%]	c	28.4	**31.6**	−59.7
	m	**16.2**	10.7	−42.2

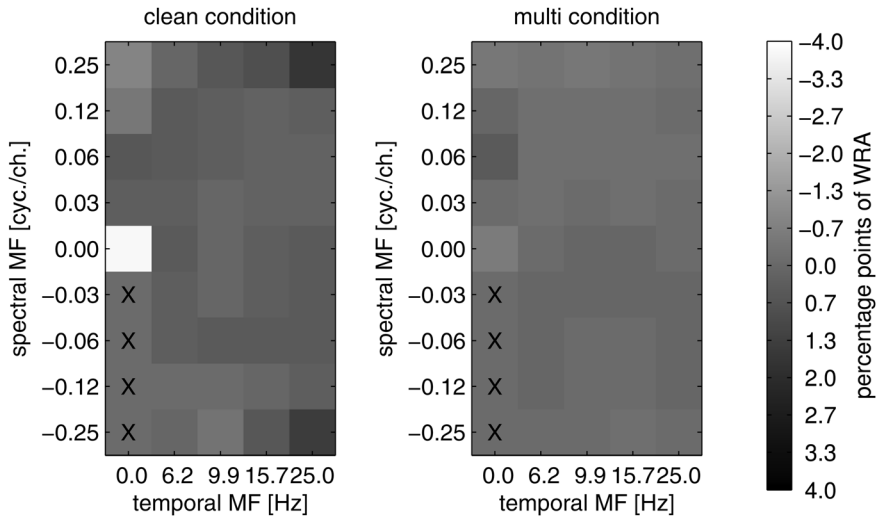

FIG. 7. Differences in overall accuracy on the Aurora 2 digit recognition task when omitting the output of filters with a particular spectro-temporal modulation frequency for multi and clean condition training. The difference in accuracy is encoded in grayscale and displayed at the position of the corresponding center modulation frequency of the omitted filter. Filters that are not used for feature extraction are marked with an X.

Sec. II B 4. The aim of this experiment is to estimate the relative importance of specific modulation frequencies. Figure 7 shows the difference between the recognition scores obtained with the original and the reduced features. Therefore, low values correspond to filters with a relatively high contribution to the recognition scores.

The patterns observed in Fig. 7 show a symmetry with respect to upward and downward filters (i.e., those with negative and positive modulation frequencies, respectively). On average, filters tuned to upward spectro-temporal patterns and filters tuned to downward spectro-temporal patterns appear to be equally important for the recognition with clean training. For multi-condition training the effect of omitting a filter is smaller (Fig. 7, right), but symmetry of upward and downward filters is not affected. The most important feature is the output of the DC filter, which encodes the level of the recording averaged over about 300 ms. The DC feature may be seen as a simple voice activity detector, and its information is not encoded in any of the other feature channels, as these do not have a DC component. Its exclusion reduces the average word recognition accuracy by about 4 percentage points, with multi-condition training it causes a drop by approximately 0.6 percentage points. The most important modulation frequency belongs to the purely spectral filter ($\omega_n = 0$ Hz) with the highest modulation frequency ($\omega_k = 0.25$ cycles/channel). It accounts for the finer spectral structure of the log Mel-spectrogram. We assume that this is the filter that best extracts information about voicing, as voicing features are represented by localized patterns that usually do not exceed two frequency channel and do not exhibit strong temporal changes.

Several filters have a detrimental effect on the overall performance, since their removal from the feature vector results in an *increase* of recognition performance: Omitting the filters with the highest modulation frequencies ($\omega_k = \pm 0.25$ cycles/channel and $\omega_n = 25$ Hz) improves the recognition performance by about 2 percentage points with clean training. The spectral filter ($\omega_n = 0$ Hz) with the lowest modulation frequency ($\omega_k = 0.03$ cycles/channel) also has a detrimental effect. This filter accounts for the very coarse spectral shape of the log Mel-spectrogram averaged over about 300 ms. It extracts mainly information about the spectral color of the communication channel. The improvement in overall scores upon deletion of specific components indicates that feature selection may further improve the recognition accuracy.

Kanedera *et al.* (1999) found that temporal modulation frequencies below 2 Hz and above 16 Hz may be detrimental for specific ASR tasks. The temporal modulation center frequencies used for the filter bank range from 6.2 Hz to 25 Hz and are subdivided into spectro-temporal upward and spectro-temporal downward filters. With GBFB features, an upper limit of about 18 Hz (cf. Table II) seems to improve performance with clean condition training from 66.2% to 69.3% overall but reduce performance with multi-condition training from 88.1% to 87.5%. The range of modulation frequencies used with GBFB features is higher than the range found by Kanedera *et al.* (1999). Some temporal modulation frequencies are only beneficial in combination with certain spectral modulation frequencies. Nemala and Elhilali (2010) found temporal modulation frequencies from 12 Hz up to 22 Hz to be useful for robust speech/non-speech recognition in an experiment that considered spectral and temporal modulation frequencies. The range of modulation frequencies used with the GBFB is in line with these findings. It is possible that an interaction between spectral and temporal modulation frequencies results in a shift of the specific frequencies important for ASR.

The most frequent temporal modulation frequency in speech is 4 Hz, but it was not found to be of particular importance for the recognition of connected digits that spectro-temporal filters tuned to 4 Hz existed at the feature level. This does not mean that it is of no importance at all, since temporal modulation frequencies below 6.2 Hz are captured by the purely spectral filters and the back-end models changes of this rate. An example for such a filter is the DC filter that changes with a temporal rate of up to about 4 Hz (cf. filter transfer function in Fig. 5) and plays an important role.

Another factor that might affect the overall recognition accuracy is the number of individual feature components associated with a spectro-temporal modulation frequency: The results of the filtering process is a spectro-temporal output with 23 frequency channels; in most cases, not all of these channels are included in the feature vector to avoid a high redundancy of feature components. The number of

selected channels ranges from 1 (for low values of ω_n and ω_k) to 23 (for high values of ω_n and ω_k). Since modulation filters are disregarded in the band-stop experiment, the number of components ranges from 288 to 310, which might have an effect on the overall performance.

III. ROBUSTNESS OF THE GABOR FILTER BANK FEATURES

In this part of the study the Gabor filter bank features are compared to several traditional feature extraction schemes in terms of robustness against extrinsic and intrinsic variability of speech. It is structured as follows: First, in Sec. III A the traditional feature extraction schemes which serve as reference are introduced. Then, in Sec. III B the experiments used for evaluation are presented. Finally, the results are presented and discussed in Sec. III C.

A. Baseline features

Standard Mel-frequency Cepstral Coeffient (MFCC) (Davis and Mermelstein, 1980) features are used as a reference. MFCCs are calculated by applying a discrete cosine transform to spectral slices of the Mel-spectrogram. The coefficients, encoding the spectral envelope of quasi-stationary speech segments, are then used as features for ASR. The Rastamat toolbox for Matlab (Ellis, 2005) is used to generate 13-dimensional MFCC features (MFCCs), which resemble the features obtained with the HTK package (Young *et al.*, 2001). Adding the first and second discrete derivative results in 39-dimensional features. As a second reference, cepstral mean subtraction (CMS), a blind deconvolution technique which Schwarz *et al.* (1993) found to improve recognition accuracy and robustness to changes of communication channel characteristics is applied to the MFCCs; these features are referred to as MFCC CMS. The baseline MFCC features on the Aurora 2 task from Pearce and Hirsch (2000) are referred to as MFCC WI007. As a third reference, 8th order Perceptual Linear Prediction (PLP) (Hermansky, 1990) features that have undergone additional modulation band pass filtering, are calculated with the Rastamat toolbox. The filtering emphasizes the relative differences between spectra, hence, these features are referred to as RASTA-PLP features (Hermansky and Morgan, 1994). RASTA-PLPs have been reported to be robust, especially in the presence of channel distortions (Hermansky and Morgan, 1994). The addition of delta and acceleration coefficients results in 27-dimensional feature vectors.

B. Experiments

In this section, the experimental setups that are employed to evaluate the robustness against extrinsic and intrinsic variability in speech are presented.

1. Effect of extrinsic factors (Aurora 2 and Numbers95)

For evaluation of robustness against extrinsic variability the Aurora 2 framework (Sec. II B 1) is used. Since several parameters of the GBFB features were optimized with the Aurora 2 framework, additional experiments are performed with a different speech corpus and a different state-of-the-art back end. The aim of this experiment is to check whether the results for GBFB features on the Aurora 2 task translate to a different ASR setup without further adaptation. The speech database chosen was NUMBERS95 (Cole *et al.*, 1995) that contains strings of spoken numbers collected over telephone connections. The data consists of zip codes and street numbers, extracted from thousands of telephone dialogues. In addition, this corpus contains data from male and female American-English speakers of different ages. Following the experimental setup from Zhao and Morgan (2008), the corpus was divided in a training set (with 3590 utterances which approximates to 3 h of data) and a testing set (1227 utterances or 1 h of data). There are two experimental conditions for the testing set; one contains all testing-set utterances in clean condition; the other contains the utterances in noise-added conditions. The noise-added test set is created using the principles delineated in the Aurora 2 task (Pearce and Hirsch, 2000) using noises of different signal-to-noise ratios from the NOISEX-92 collection (Varga and Steeneken, 1993).

Features were mean and variance normalized and used to train the GMM/HMM recognizer *Decipher* developed by Stanford Research International (SRI). This state-of-the-art system is used to compare spectro-temporal and other features against a competitive baseline. Gender-independent, within-word triphone HMM models were based on a phone model comprising 56 consonants and vowels. Parameters were shared across 150 states clustered with a phonetic decision tree, and a diagonal-covariance GMM with 16 mixture components modeled the observation distribution. Maximum Likelihood estimation was used to estimate the parameters. Features are used either as direct input to Decipher, or processed in a Tandem system (Hermansky *et al.*, 2000) that uses a multi-layer perceptron (MLP) to estimate the phone posterior probabilities for each feature frame. The posteriors are then log-transformed and decorrelated with a principal component analysis, in order to match the orthogonality assumption of the HMM decoder. For experiments that employ MLP-processing, the training of the neural net was carried out with phonetically labeled digit sequences from Numbers95 training set. The phoneme labels were obtained from forced alignment. The MLP used 9 frames of temporal context which resulted in $9 \times 331 = 2927$ input units, 160 and 56 units were used for the hidden and output layer, respectively. For the last set of experiments, 13-dimensional MFCC features with delta and double-delta features were appended to the MLP-transformed Gabor features, resulting in 71-dimensional feature vectors, since this has been reported to increase accuracies in other research that used spectro-temporal features as input to ASR (Zhao and Morgan, 2008). The results for the MFCC, MFCC CMS, RASTA-PLP and GBFB features are presented, compared and discussed in Sec. III C 1 on the Aurora 2 task, and in Sec. III C 2 on the Numbers95 task.

2. Effect of intrinsic factors (OLLO framework)

To evaluate the robustness against *intrinsic* variability in speech, an experimental framework that aims at the

analysis of factors such as speaking style, effort, and rate is proposed. In this framework the sensitivity of different feature types against such variabilities is evaluated by performing experiments with a mismatch between the training and test data. The degradation in performance quantifies the robustness against a specific mismatch. A statistical test, McNemar's Test as suggested by Gillick and Cox (1989), is employed to test the results for significant differences between the feature types.

The speech database used for this framework is the Oldenburg Logatome Corpus (OLLO) (Wesker *et al.,* 2005), which consists of nonsense vowel-consonant-vowel (VCV) and consonant-vowel-consonant (CVC) logatomes with identical outer phonemes (e.g., [p u p] or [a p a]). The database contains 150 different logatomes (70 VCVs and 80 CVCs), spoken by German speakers in different speaking styles. During the recordings, the speakers were asked to produce the utterances normally, with varied speaking effort (loud and soft speaking style), varied speaking rate (fast and slow), and with rising pitch, which is referred to as category "questioning." Three repetitions of each logatome in each speaking style were collected in order to obtain a sufficient amount of ASR training data. This resulted in 150 (logatomes) $\times$ 6 (speaking styles) $\times$ 3 (repetitions) = 2700 utterances per speaker.

For the OLLO framework that we propose to evaluate robustness against intrinsic variability in speech, speech data from ten speakers without dialect is used. Six training and six test conditions are defined, which correspond to the various speaking styles contained in the OLLO corpus (fast, slow, loud, soft, questioning and normal). Training and testing on each condition resulted in 36 individual experiments. The experiments are carried out using a 10-fold cross validation, i.e., speech signals of nine speakers are used for training, and the data of the remaining speaker is used for testing. This procedure is repeated for all speakers, and the individual scores are averaged.

As for the Aurora 2 framework, results for MFCC features serve as baseline. These are fed to an HMM using HTK (Young *et al.,* 2001). The HMM is configured as word recognizer, i.e., the classification task is to make a 1-out-of-150 decision based on a dictionary that contains the transcription of the 150 logatomes. The number of HMM states per logatome is set to 16, which was found to be the optimal value in pilot experiments for MFCC features. Other parameters, such as the increase of Gaussian mixtures during training, are copied from the Aurora 2 setup. Additionally, performance of MFCC features with CMS and RASTA-PLP features is evaluated. Since the PLP part of this algorithm accounts for the reduction of speaker-dependent information it is interesting to see whether it improves the robustness against intrinsic variability. The results are presented in Sec. III C 3.

C. Results and discussion

1. Robustness against extrinsic variability (Aurora 2)

This section presents the results of recognition experiments with GBFB, MFCC, MFCC CMS and RASTA-PLP features that are carried out with the aim of quantifying the robustness against extrinsic variability (additive noise and channel distortions) on the Aurora 2 task (employing the HTK recognizer).

Absolute results for the various feature types are presented in Table V. In terms of average word recognition accuracies (WRAs) GBFB features outperform MFCC and RASTA- PLP features with clean (multi) condition training by 8 (1) percentage points and 2 (3) percentage points, respectively. With cepstral mean subtraction MFCC features achieve a slightly higher average WRA than GBFB features. The overall relative improvement over MFCC WI007 standard features, which is calculated as described in Sec. II B 1 is presented in Table VI. GBFB features improve the WER of standard MFCC WI007 features by more than 16% on average with multi condition training and by 28% on average with clean condition training. The use of MFCCs with CMS improves the baseline by 12% on average with multi condition training and by 27% on average with clean condition training. When concatenating GBFB features with MFCC features from the Rastamat toolbox with CMS applied, a further improvement of a few percent is achieved, indicating that these feature types carry complementary information. RASTA-PLP features outperform the standard MFCC WI007 features by about 14% with clean condition training, but with multi condition training they perform 31% worse than MFCCs.

The relative improvements over the baseline for the test sets A, B, and C are also presented in Table VI. In addition to the reference features, the performance of GBFB features concatenated with MFCC CMS features is shown. GBFB features outperform the MFCC WI007 baseline in all test conditions (test sets A, B and C). For multi-condition training, the relative improvements for test set A and test set B are comparable with improvements of about 13%, which indicates that GBFB features generalize as well as MFCC features with respect to mismatches in noise types when training with noisy data. For test set C, the relative improvement for GBFB features is more distinctive with about 28%. MFCC features with CMS improve the WI007 baseline in test set B and C, i.e., when noise or communication channel characteristics changed compared to the training data. The improvements with test set C (channel distortions) for MFCCs with CMS is smaller than with GBFB features. RASTA-PLP features perform worse than the MFCC WI007 baseline with multi condition training on all test sets. For test set C (channel distortions), the difference between RASTA-PLP and MFCC features is smaller than on test set A and B.

For clean training, the differences between GBFB and MFCC WI007 features are larger compared to multi-condition training with a relative decrease of the WER for GBFB features in the range of 15% to 40%. With MFCC CMS features and clean condition training the improvements are also larger compared to the multi condition training. The smaller relative improvements in test set C are a result of the relatively high performance of the MFCCs (cf. Table V). RASTA-PLP are consistently better than the baseline for clean condition training, but do not improve results with multi-condition training.

TABLE V. Recognition accuracies in percent for GBFB, MFCC WI007, MFCC CMS, and RASTA-PLP features on the Aurora 2 task for different noise conditions, average word recognition accuracies for each test set and standard deviation over all noise conditions. The average values presented here are obtained by averaging over SNRs from 0 dB to 20 dB.

		Test set A				Test set B				Test set C			
		Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Subway m	Street m	Average	rms
GBFB	multi	89.0	88.0	86.1	88.1	90.0	88.2	90.7	85.9	88.8	86.3	88.1	**1.62**
	average			87.8			88.7				87.6		
	clean	70.9	67.0	60.0	64.3	69.2	64.5	68.9	65.0	68.8	63.4	66.2	3.33
	average			**65.6**			66.9				**66.1**		
RPLP	multi	87.5	84.6	83.7	83.7	83.8	84.5	85.6	82.1	87.1	84.3	84.7	1.64
	average			84.9			84.0				85.7		
	clean	64.3	63.1	59.5	59.9	66.3	63.8	66.4	63.1	64.5	63.7	63.5	**2.30**
	average			61.7			64.9				64.1		
MFCC	multi	89.1	88.4	86.8	80.0	87.8	88.3	86.2	85.7	87.0		1.67	
	average			88.1			87.2				84.6		
	clean	66.7	47.8	58.1	62.3	50.0	60.7	49.6	53.1	65.3	66.7	58.1	7.40
	average			58.7			53.4				66.0		
$MFCC^{CMS}$	multi	90.3	89.8	84.9	88.0	90.0	87.9	91.1	86.7	89.9	87.9	**88.7**	1.92
	average			**88.2**			**88.9**				**88.9**		
	clean	64.1	67.7	62.2	62.8	71.3	65.6	72.0	67.7	64.2	65.4	**66.3**	3.35
	average			64.2			**69.2**				64.7		

The standard deviation of recognition scores for various noise conditions is reported as a measure for the stability of scores in the last column of Table V. The results for clean condition training are of special interest in this case, since they can be interpreted as the robustness in the presence of unknown noise sources. The standard deviation for MFCCs (7.4 percentage points) is approximately twice as high as for GBFB and three times as high as for RASTA-PLP features (3.4 percentage points and 2.3 percentage points, respectively). This indicates that GBFB and RASTA-PLP features are less sensitive to mismatches between training and test data than MFCC features. When applying CMS to the MFCC features the standard deviation decreases to the level of GBFB features. For multi-condition training, the standard deviations are smaller than 2 percentage points, with only small differences between the feature types but MFCCs with CMS, which show a slightly higher standard deviation.

TABLE VI. Relative reduction of the word error rate obtained with GBFB, MFCC CMS, RASTA-PLP features, and with GBFB features concatenated with MFCC CMS features compared to the MFCC WI007 baseline for the test sets A, B and C.

		Test set A 4 conditions	Test set B 4 conditions	Test set C 2 conditions	Average over all conditions
GBFB	clean	22.9	40.6	15.1	28.4
	multi	11.4	15.2	27.7	16.1
$MFCC^{CMS}$	clean	15.9	**45.5**	10.9	26.7
	multi	3.6	16.1	19.0	11.6
RPLP	clean	2.6	31.4	4.0	14.4
	multi	−33.3	−40.0	−12.7	−31.9
GBFB and $MFCC^{CMS}$	clean	**26.3**	43.8	**18.0**	**31.6**
	multi	**16.7**	**24.4**	**33.0**	**23.0**

A comparison of the relative improvements of GBFB features over MFCC features in Table VI with the absolute results in Table V shows that the differences between both in terms of WRAs is rather small. This suggests that the improvements of GBFB over MFCC features are obtained at high SNRs. This is investigated further by separating the WRA results by SNRs. The average WRAs for each feature type and for each SNR are depicted in Fig. 8. The ordinate is scaled as a logarithmic error axis and labeled with the corresponding WRA. The distance of two horizontal lines corresponds to a halving/doubling the WER so that the results in terms of relative improvements are projected linearly, i.e., they are proportional to the relative improvement of the averages over all noise conditions.

For all feature types, a strong decrease of the WRA is observed when the noise level is raised, with 95% WRA for clean utterances to down to scores below 30% at an SNR of −5 dB. The major differences between the feature types are observed at high SNRs (20 dB to 5 dB). For clean training, the decrease in performance is more pronounced than for multi-condition training. While using noisy training data and testing with clean utterances results in lower scores compared to clean training, the overall performance (tested over multiple SNRs and noise types) is improved with multi-condition training as expected. When training with clean utterances, RASTA-PLP features outperform the MFCC WI007 baseline at almost all SNRs, which confirms the observation that RASTA-PLPs are more robust than short-term spectrum based features in unknown noise conditions (Hermansky and Morgan, 1994). However, for multi-condition training, which allows the ASR system to adapt to different noise types, the MFCCs produce higher scores than RASTA-PLPs. GBFB features improve the scores of MFCC WI007 and RASTA features at almost all SNRs: The robustness

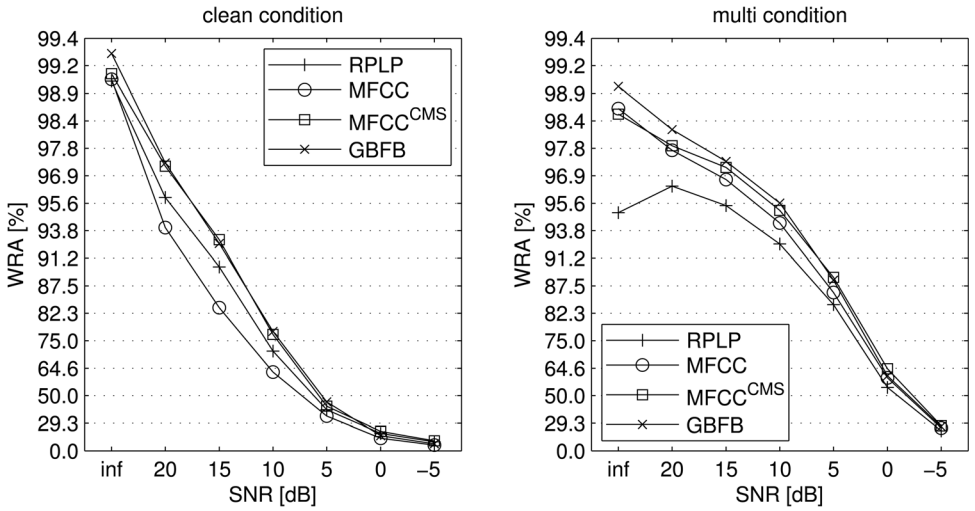

FIG. 8. Recognition accuracies in percent for GBFB, MFCC and RASTA-PLP features at different test SNRs for multi and clean condition training on the Aurora 2 task. The ordinate is a logarithmically scaled WER-axis and labeled with the corresponding WRA. The distance of two horizontal lines corresponds halving/doubling the WER.

against additive noise is found to be higher than for RASTA features over a wide range of test conditions (i.e., clean signals and SNRs from 5 dB to 20 dB), and additionally are found to outperform the MFCC WI007 baseline for multi condition training in these test conditions. MFCCs with CMS and GBFB features perform similarly well. However, when testing on clean or low-noise signals with multi-condition training, GBFBs outperform all feature types.

The average relative improvement of GBFB, MFCC CMS and RASTA-PLP over MFCC WI007 features depending on the SNR (averaged over all noise conditions) is depicted in Table VII. The results are comparable to those presented in Fig. 8. GBFB features outperform MFCCs at all SNRs. The best improvements are obtained at SNRs above 0 dB SNR, while at low SNRs the differences are negligible. While RASTA-PLP features outperform MFCCs with clean training at all SNRs they do not improve MFCC results when testing on clean data. The relative improvements for low SNRs (0 dB, −5 dB) are higher than with GBFB features but still below 6%. This means that RASTA-PLP and GBFB features are more robust than MFCCs when the noise signal energy still is about 5 dB below the level of the speech signal energy. When learning the noise characteristics (multi-condition), GBFB features perform better and RASTA-PLP features perform worse than MFCCs, with the greatest differences in relative improvements at high SNRs. MFCCs with CMS improve the MFCC WI007 baseline at almost all SNRs, with the single exception of multi-condition training and clean testing. When testing on clean data GBFB features improve the MFCC baseline by more than 6%, while the baseline was not improved with the other feature types. For testing on clean speech data, GBFB features improve the baseline by about 25%.

A 28% relative improvement of GBFB features over the MFCC baseline is observed when the channel characteristics of training and testing differ. We assume that MFCCs are stronger affected by such influences since the spectrogram is integrated over the full bandwidth, which might be a disadvantage compared to the localized GBFBs. Cepstral mean subtraction seems to alleviate this disadvantage, but not to the extent that was observed for GBFB features. Further, considering even higher frequencies (above 4 kHz) could be beneficial with the Gabor filter bank features. While the MFCC features would change fundamentally, for the Gabor filter bank features it would mean an extension to more center frequencies. This should be evaluated on a suitable task in the future.

GBFB features were shown to perform better in the high SNR range from 20 dB to 5 dB than MFCC and RASTA-PLP features and equally well at lower SNRs (0 dB and −5 dB) on the Aurora 2 task, which evaluates robustness of ASR systems against extrinsic variability. GBFB features also slightly outperform MFCC features with CMS. This suggests that the physiologically inspired representation of speech signals by GBFB features is more robust to extrinsic variability than those of MFCCs and RASTA-PLPs over a wide range of SNRs and is similarly robust to extrinsic variability as MFCC with CMS. Further, improvements of about 25% for testing on clean data are observed which points out the beneficial effect of spectro-temporal information on feature level.

TABLE VII. Relative reduction of the word error rate obtained with GBFB and RASTA-PLP features compared to the MFCC baseline (averaged over all noise conditions). Values in the column *average* are averaged over SNRs from 20 dB to 0 dB.

SNR [dB]		∞	20	15	10	5	0	−5	Average
GBFB	multi	23.9	23.4	21.0	21.3	13.5	1.7	1.8	16.2
	clean	27.0	45.1	45.1	34.9	14.3	2.8	1.5	28.4
RASTA-PLP	multi	−274.6	−56.4	−38.5	−31.8	−18.4	−14.5	−3.7	−31.9
	clean	−3.8	16.9	28.4	16.7	4.7	5.3	3.9	14.4
MFCC^{CMS}	multi	−7.6	5.3	13.4	13.0	16.3	10.1	2.9	11.6
	clean	5.5	42.2	45.9	29.85	8.00	7.6	5.2	26.7

2. Robustness against extrinsic variability (Numbers95)

In this section the results for the NUMBERS recognition task, which was conducted with the aim of checking whether the results from the Aurora 2 task translate to a different ASR setup, are presented. Absolute and relative results obtained on the NUMBERS recognition task with the SRI Decipher recognizer are shown in Table VIII. In this scenario, MFCC and MFCC CMS features perform best, while for GBFB and RASTA-PLP features relatively high error rates are observed. A possible reason may be that GBFB features encode up to 400 ms context and RASTA-PLP features up to 200 ms context, and may thus not be suited as well as the MFCCs (up to 100 ms context) for triphone based models.

We then tested if mapping the features to phoneme posteriors, which we assume to be suitable to build phone base models, by means of a multi-layer perceptron (MLP) improves the recognition performance. This MLP processing, which was reported to improve results in earlier studies (Hermansky *et al.*, 2000), almost halved the error rate of GBFB features without MLP processing for clean testing and also improved the results with RASTA-PLP features, but the performance was still below the baseline. Using MFCCs and MFCCs with CMS in conjunction with MLP processing leads to small improvements when testing on noisy data, but not for testing on clean data. The results with "long-term context" features, i.e., GBFB features and also RASTA-PLP to a smaller extent, improved much more by the MLP processing than the results with the already well performing "short-term context" features. Another reason for the high error rate with GBFB features may be the high dimensionality of the features. While the GMM/HTK back end of the Aurora 2 framework had no problems with high dimensional features, the Decipher recognizer may be tuned to the dimensionality of typical feature types, hence performing better with the low dimensional MLP processed GBFB features.

The improvements over the MFCC baseline that were observed on the Aurora 2 task with GBFB features do not translate directly to setups with different back ends. This is because the back end imposes strong restrictions upon the statistical characteristics of the used features. These restrictions depend on many factors like the training material, the acoustic model type, and the complexity of the recognizer. We assume that the shorter triphone models of the Decipher back end favor features with less temporal context compared to the whole word models of the HTK recognizer on the Aurora 2 task. However, adapting the features to the restrictions of the back end improves the recognition performance. GBFB features are long-term context features and seem to work better with models that can make full use of long-term context (word models).

The fact that error rates were lower when combining MFCC and GBFB features for the Aurora 2 task motivated a combination of MLP-processed features with MFCCs. For this setup, the MFCC baseline is outperformed by more than 10% (for combinations with MFCCs), and 14–15% for combinations with MFCC CMS. We also tested other combinations (such as MLP-processed *spectral* features that are combined MFCCs); however, none of these yielded results above the baseline.

With the Decipher back end, the baseline was not improved when only using GBFB features, but when using the features in a Tandem system and combining them with spectral features, the baseline was outperformed by 14–15%. This result confirms earlier studies that reported an increase of the robustness of ASR system against additive noise and channel distortions when using MLP-processed spectro-temporal features in conjunction with concatenated MFCCs (Meyer and Kollmeier, 2011a; Zhao *et al.*, 2009). It also supports the hypothesis that MFCCs and GBFB features encode complementary information that is useful for robust ASR.

TABLE VIII. Word error rates for the NUMBERS95 task with SRI's ASR system Decipher. Features were either used as direct input to the classifier, processed with an MLP, or first MLP-processed and then concatenated with a different feature vector

		Absolute WER		Rel. imp.	
	Feat. dim.	Clean	Avg. noisy	Clean	Avg. noisy
MFCC	39	3.7	19.4	–	–
$MFCC_{CMS}$	39	3.7	17.8	1.1	8.1
RASTA-PLP	27	6.0	23.2	−59.6	−19.7
GBFB	311	9.1	22.9	−142.3	−16.2
MLP (MFCC)	32	4.0	19.2	−7.9	1.0
MLP ($MFCC_{CMS}$)	32	3.7	16.9	1.1	12.7
MLP (RASTA−PLP)	32	5.7	20.8	−51.1	−7.6
MLP (GBFB)	32	4.6	19.9	−21.9	−2.6
MLP (GBFB) and MFCC	71	3.3	16.8	10.7	13.5
MLP (GBFB) and $MFCC_{CMS}$	71	3.2	16.6	15.2	14.1

3. Robustness against intrinsic variability

This section presents the results of recognition experiments with GBFB and baseline features features that are carried out with the aim of quantifying the robustness against variability due to intrinsic sources (arising from variation in speaking rate, effort and style). The ASR task is to classify VCV and CVC utterances from the OLLO database, as described in Sec. III B 2. The absolute word recognition accuracies are depicted in Table IX. Scores are presented for each combination of training and test speaking styles, which results in 6×6 individual scores per feature type. RASTA-PLP and MFCC CMS features produce almost consistently worse scores than MFCC and GBFB features and are therefore not included in Table IX.

When averaging over all scores obtained for mismatched training and test conditions (off-diagonal elements in Table IX), the recognition scores for GBFB and MFCC features are very similar with 59.3% and 58.8%. RASTA-PLPs produce an average of 55.6% (not shown) and MFCC CMS features produce an average of 56.4% (also not shown). All feature types exhibit similar error patterns, which are depicted in Fig. 9. Not surprisingly, the best scores are obtained with matched condition training. Compared to

TABLE IX. Absolute WRA in percent for GBFB and MFCC features on the OLLO logatome recognition task. Averages are calculated over mismatched conditions. Matched conditions are printed in italics and are not considered for averages.

	GBFB							MFCC						
Train | Test	Fast	Slow	Loud	Soft	Quest.	Normal	Average	Fast	Slow	Loud	Soft	Quest.	Normal	Average
Fast	*74.0*	47.0	62.7	52.6	49.5	72.7	56.9	*72.4*	52.2	60.8	49.7	49.4	72.3	56.9
Slow	45.6	*76.7*	46.5	66.3	39.4	69.9	53.5	51.3	*77.5*	50.5	66.2	50.0	73.1	58.2
Loud	70.3	56.3	*78.7*	47.7	50.5	75.6	60.1	68.6	58.9	*77.1*	43.1	52.5	72.7	59.2
Soft	51.9	64.0	42.9	*74.1*	49.1	67.7	55.1	50.1	62.7	31.9	*71.0*	45.0	64.1	50.8
Quest.	61.8	65.7	56.0	68.0	*78.3*	74.9	65.3	61.0	66.0	54.8	63.9	*76.8*	72.2	63.6
Normal	70.7	65.7	64.4	66.5	56.0	*81.4*	64.7	70.6	68.4	61.3	64.4	56.4	*79.9*	64.2
Average	60.1	59.7	54.5	60.2	48.9	72.2	59.3	60.3	61.6	51.8	57.5	50.6	70.9	58.8

matched train-test conditions, the word recognition accuracies of the mismatched conditions show a degradation of about 17 percentage points on average. For MFCCs, the category "normal" for training yields the highest scores when considering the average over all six test conditions. For GBFB features, the category "questioning" for training yields slightly better (0.6 percentage points) word recognition accuracies than the category "normal." When using normally spoken utterances for the training, the reduction in WER is roughly −70% when testing on mismatch conditions (average over all feature types).

For the chosen order of sources of variability in Fig. 9, a checker board pattern is observed in the upper left part of each matrix. Relatively high accuracies are obtained for the training-test pairs (fast, loud) and (slow, soft), which indicates that utterances from these categories share properties that are embedded in the acoustic models of the HMM during training. On the other hand, the pairs (fast, slow), (fast, soft), (loud, soft) and (loud, slow) yield a score that is degraded by about 24 percentage points on average (average over all feature types) compared to the respective matched condition.

While GBFB and MFCC features perform similarly well on average, a detailed analysis of the recognition results with respect to speaking rate, style, and effort reveals systematic differences. Figure 10 shows in which particular conditions the differences between MFCC features and GBFB features are significant, i.e., the p-values are less than 0.01 according to McNemar's Test as proposed by Gillick and Cox (1989). The differences in terms of relative improvement of WER are depicted in Table X. Only the mismatch conditions (off-diagonal elements) are considered for the average. These values can be interpreted as the sensitivity of GBFB features (compared to MFCC results), or the robustness against intrinsic variability.

The results show that on average GBFB features are slightly more sensitive against such mismatches (with a 0.2% relative degradation when averaging over all combinations of training and testing). The relative reduction of the WER with GBFB features compared to MFCCs shows that MFCCs exhibit a better recognition performance for high and low speaking rate (categories "fast" and "slow"), while GBFB features are better suited when the talker changes his speaking effort (categories "loud" and "soft"). This trend is consistent both for training and for testing. Interestingly, when the recognizer is *trained* with utterances with rising pitch ("questioning"), GBFB feature perform better than MFCC features (row "questioning" in Table IX). On the other hand, when *testing* is performed with logatomes spoken as question, this results in higher scores with MFCC features than with GBFB features (column "questioning" in Table X).

On average, the performance with MFCC features deteriorates by about 70% (relative improvement calculated as explained in Sec. II B 1), the GBFB features' performance drops by about 80% when training and test data categories mismatch. Compared to MFCCs, GBFB features seem to perform similarly well on average in the tested mismatching conditions of intrinsic sources of variability. However, they appear to be slightly more susceptible to such variations than MFCCs, since they tend to perform better in matched conditions, which are not considered for averages.

In the presence of intrinsic variation (measured with the OLLO recognition task, cf. Sec. III B 2) considerable degradations are observed for all feature types. Compared to the matched condition scores, the average relative increase of the word error rate is between 70 and 85% (for MFCC and MFCC CMS features, respectively). In order to analyze the robustness against intrinsic factors, the scores obtained with mismatched training are of special interest. In the presence

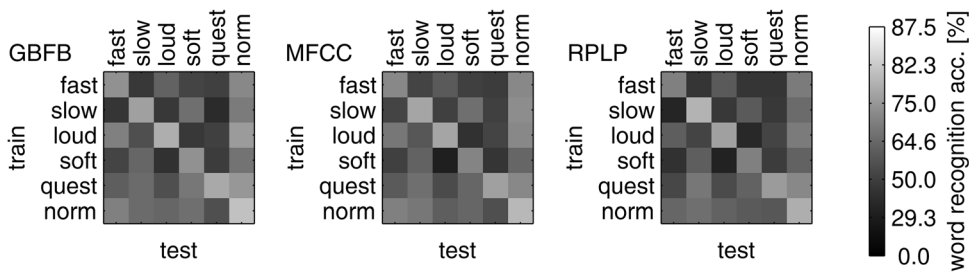

FIG. 9. Logarithmic word error rates for different training and testing conditions on the OLLO logatome recognition task. The colorbar indicates the corresponding word recognition accuracies. (left) GBFB features; (middle) MFCC features; (right) RASTA-PLP features.

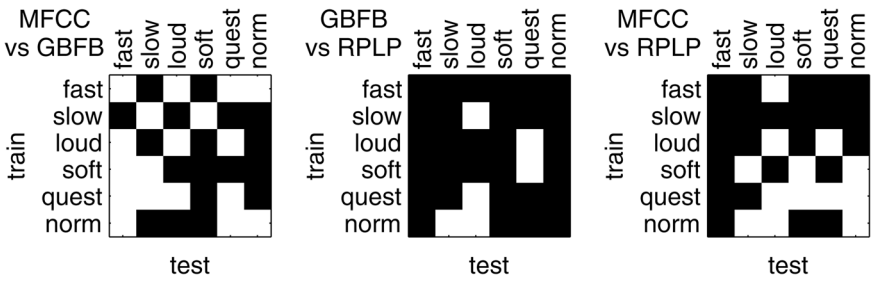

FIG. 10. Analysis of differences between the feature types according to McNemar's Test. Black: Significant differences with $p < 0.01$; white: Not significant. (left) Differences between GBFB and MFCC features; (middle) differences between GBFB and RASTA-PLP features; (right) Differences between MFCC and RASTA-PLP features.

of variation caused by intrinsic sources, GBFB and MFCC features exhibit a comparable overall performance. However, when individual sources of variability are considered, the error patterns for both feature types show statistically significant differences, indicating that these feature types carry—at least to some extent—complementary information.

Using mismatched conditions in training and testing shows that the training-test pairs (fast, loud) and (slow, soft) produce relatively high accuracies. This trend is observed for all feature types. The combinations (fast, soft) and (slow, loud) on the other hand produce rather low scores. It may be that these categories share several acoustic properties, since speakers, e.g., unconsciously increase their speaking effort when asked to produce an utterance with high speaking rate. Such an interaction might also explain the high scores for the pair (soft, slow).

GBFB features are observed to perform better than MFCCs when *training* on utterances pronounced with rising pitch (category "questioning"), but worse when *testing* on utterances of this category. A possible explanation for this observation is that GBFB features can account for spectral details such as pitch information. However, to account for the larger variability caused by changes in pitch, the according speech data has to be included in the training material.

For matched training and test conditions, the best average results are obtained with GBFB features. However, GBFB features are not found to be more robust than MFCC features against intrinsic variability, i.e., spectro-temporal information does not seem to improve robustness against intrinsic variability in general. The differences observed between the feature types indicate that the information captured in the feature calculation process is at least partially complementary; hence, the combination of these features (e.g., in a multi-stream framework) could result in an improvement of the ASR performance.

TABLE X. Relative improvement of GBFB features over MFCC features in percent. Scores for matched conditions (diagonal elements of the table, printed in italics) are not considered for the average values.

Train | Test	Fast	Slow	Loud	Soft	Questioning	Normal	Average
Fast	*+6*	−11	+5	+6	+0	+2	+0
Slow	−12	*−4*	−8	+0	−21	−12.0	−11
Loud	+5	−6	*+7*	+8	−4	+11	+3
Soft	+4	+4	+16	*+11*	+8	+10	+8
Quest.	+2	−1	+3	+11	*+6*	+10	+5
Normal	+0	−9	+8	+6	−1	*+8*	+1
Average	−0	−5	+5	+6	−4	+4	+1

For RASTA-PLP and MFCC CMS features, relatively low scores are obtained. The fact that both the training and testing with the OLLO database are performed with clean utterances might explain this observation for RASTA-PLP, since the Aurora 2 experiment showed that these features only improved the baseline for additive noise and channel distortions. Moreover, the calculation of RASTA-PLPs includes temporal filtering, which might be suboptimal for very short utterances such as the phoneme combinations used for the OLLO corpus, although GBFB features also capture temporal information to a comparable extent. For CMS, an integration over the whole utterance is needed. Maybe the shortness of the utterances does not allow for a good estimation of the mean value, thus resulting in a mismatch that deteriorates performance.

IV. SUMMARY AND FURTHER DISCUSSION

A. Robustness of GBFB features against extrinsic variability

The performance of a robust speech recognition system depends on the interaction of its parts. The results presented in this study show that improvements over a MFCC baseline can be obtained with physiologically inspired spectro-temporal features when the back end's assumptions about the statistical feature characteristics are met. It can be assumed that the properties of the Gabor filter bank result in a filter output with limited redundancy between individual components and mostly independent features with up to 400 ms of temporal context. Depending on the task and the back end it may be favorable to apply MLP processing to the GBFB features in order to meet the back end's assumptions about the features. In this case improvements over the unprocessed GBFB features can be expected, but not necessarily an improvement of a MFCC baseline.

B. Complementary information

The experiments show that the combination of MFCC CMS and GBFB features, possibly processed with an MLP, results in a further increase of recognition performance. Presumably, there most possibly is a part of information important for ASR represented in a better suited form by MFCCs than by GBFB features and vice versa. This also means that neither MFCC nor GBFB features are sufficient to extract all the characteristics of human speech.

Earlier studies using spectro-temporal features for ASR presented evidence that MFCCs and spectro-temporal features carry complementary information (Meyer and Kollmeier,

2011a; Zhao *et al.,* 2009). This finding is also supported by the experiment that analyzed the sensitivity against intrinsic variation, since the performance obtained with MFCC and GBFB features significantly differ in many conditions. For example, cepstral features are found to be better suited for recognition of fast and slowly spoken utterances, while GBFB features produce better results when the speaking effort is varied. The results of the multi-stream experiment carried out on the Aurora 2 task, which improves performance over GBFB and MFCC CMS features by concatenating them also supports this finding.

C. Future work

The Gabor filter bank could be used for speech analysis in order to evaluate the importance of modulation frequencies: The integration of the outputs of the localized Gabor filters results in a spectro-temporal representation resembling the original spectrogram. When isolated spectro-temporal components are removed from the filter bank, their contribution to speech recognition may be assessed in tests with human listeners (resembling the ASR band-stop experiments carried out in this study).

The results investigating intrinsic variation of speech show that spectro-temporal and purely spectral ASR features produce significantly different results depending on the specific source of variability. Further, small improvements over using GBFB features are achieved when combining them with MFCC CMS features. Based on these observations, it may be worthwhile to further investigate methods to combine information from different feature streams, thereby exploiting the complementary information of the feature types. The output of the Gabor filter bank also contains purely spectral output, which may not be required (or even detrimental) when combined with MFCC features, which may also be subject of future investigations. Alternatively, the purely spectral output of the GBFB might be modified to closely resemble the extraction of cepstral features, which would effectively integrate the informational content of MFCCs into Gabor features.

The GBFB features extend naturally to higher frequency bands. It should be evaluated if this behavior has an advantage over MFCC features that always project the whole bandwidth of the log Mel-spectrogram.

The parameters of the Gabor filter bank (i.e., the optimal number of oscillations under the envelope) are optimized on the Aurora 2 digit recognition task, but also show good performance on the OLLO logatome recognition task. However, when changing the back end, the GBFB features do not necessarily meet the assumptions made about them and can perform worse than traditional features. In this case the robustness of these systems may be improved by processing the GBFB features with a MLP and concatenating MFCC features. This suggests that the proposed GBFB features, possibly with MLP processing, may be applicable to a wider range of ASR recognition tasks with the same parameters, which should be assessed in future experiments. To further validate the findings, GBFB features should be tested on a large vocabulary speech recognition task.

It seems that not all of the 311 filter outputs extract useful information. Especially the highest spectro-temporal modulation frequencies seem to have a negative effect on the recognition performance. It could be that covering a rectangular region of the modulation domain is not optimal. Hence, feature selection techniques could further improve the performance.

V. CONCLUSIONS

The most important findings of this work can be summarized as follows.

(1) The use of spectro-temporal Gabor filter bank (GBFB) features increases the robustness of ASR systems against additive noise and mismatches of channel transmission characteristics (i.e., *extrinsic* sources of variability) compared to MFCC and RASTA-PLP features. For this, it can be necessary to process the GBFB features with a multi-layer perceptron (MLP) and combine them with MFCCs depending on the task and the back end. A MFCC baseline was also improved for high SNRs and clean speech. With a standard GMM/HMM recognizer, improvements of over 40% with clean training and over 20% with multi training were observed when the GBFB features were used as direct input to the classifier. A state-of-the-art baseline system was outperformed by 14–15% when GBFB features were first processed with a MLP and then combined with MFCC features. These findings indicate that the proposed feature extraction scheme results in a good representation of speech signals for ASR tasks. GBFB and MFCC features were found to extract partly complementary information regarding extrinsic and intrinsic sources of variability, which may be exploited in feature stream experiments.

(2) On average, MFCC and GBFB features are similarly affected by *intrinsic* variability of speech, while for RASTA-PLP features and MFCCs with CMS higher degradations are observed. When analyzing train-test pairs with unmatched intrinsic variations, the MFCC and GBFB scores show significant differences, which shows that the feature types exhibit different strength and weaknesses with respect to intrinsic factors.

(3) The analysis of specific modulation frequencies for ASR with GBFB features shows that temporal modulation frequencies from 6 Hz to 25 Hz and spectral modulation frequencies from 0.03 cycles/channel to 0.25 cycles/channel are important for robust speech recognition. Besides the information about the input level, spectral modulation frequencies of about 0.25 cycles/channel were found to be especially important for robust speech recognition. When using spectro-temporal features for ASR, the usable temporal modulation frequencies are shifted to higher frequencies than reported in the literature that analyzed spectral and temporal information separately.

ACKNOWLEDGMENTS

Supported by the DFG (SFB/TRR 31 "The active auditory system"). Bernd T. Meyer's work is supported by a

post-doctoral fellow-ship of the German Academic Exchange Service (DAAD). We would like to thank Jörg-Hendrik Bach and Jörn Anemüller for their support and contribution to this work. Thanks also to Suman Ravuri and Andreas Stolcke for providing support for the experiments with SRI's recognition system, and the two anonymous reviewers, who made valuable suggestions to improve this study.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. (**2007**). "Automatic speech recognition and speech variability: A review," Speech Commun. **49**, 763–786.

Bouvrie, J., Ezzat, T., and Poggio, T. (**2008**). "Localized spectro-temporal cepstral analysis of speech", in *Proceedings of ICASSP 2008*, pp. 4733–4736.

Chi, T., Ru, P., and Shamma, S. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. **118**, 887.

Cole, R. A., Noel, M., Lander, T., and Durham, T. (**1995**). "New telephone speech corpora at CSLU," in *Proceedings of Eurospeech 1995*, p. 95.

Cooke, M., and Scharenborg, O. (**2008**). "The Interspeech 2008 consonant challenge," in *Proceedings of Interspeech 2008*, pp. 1781–1784.

Davis, S., and Mermelstein, P. (**1980**). "Comparison of parametric representations for mono-syllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech Signal Process. **28**, 357–366.

Domont, X., Heckmann, M., Joublin, F., and Goerick, C. (**2008**). "Hierarchical spectro-temporal features for robust speech recognition," in *Proceedings of ICASSP 2008*, pp. 4417–4420.

Ellis, D. (**2005**). "PLP and RASTA (and MFCC, and inversion) in MATLAB," available at http://www.ee.columbia.edu/dpwe/resources/matlab/rastamat/ (Last visited February 27, 2012).

ETSI Standard 201 108 v1.1.3 (**2003**). It is available at the ETSI website: http://www.etsi.org/WebSite/Technologies/DistributedSpeechRecognition.aspx.

Ezzat, T., Bouvrie, J., and Poggio, T. (**2007a**). "AM-FM demodulation of spectrograms using localized 2D max-Gabor analysis," in *Proceedings of ICASSP 2007*, Vol. **4**, pp. 1061–1064.

Ezzat, T., Bouvrie, J., and Poggio, T. (**2007b**). "Spectro-temporal analysis of speech using 2-D gabor filters," in *Proceedings of Interspeech 2007*, pp. 506–509.

Gillick, L., and Cox, S. (**1989**). "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings of ICASSP 1989*, Vol. 1, pp. 532–535.

Gramss, T. (**1991**). "Word recognition with the feature finding neural network (FFNN)," in *Proceedings of the International Workshop Neural Networks Signal Process*., pp. 289–298.

Heckmann, M., Domont, X., Joublin, F., and Goerick, C. (**2008**). "A closer look on hierarchical spectro-temporal features (HIST)," in *Proceedings of Interspeech 2008*, pp. 894–897.

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

Hermansky, H., Ellis, D. P. W., and Sharma, S. (**2000**). "Tandem connectionist feature extraction for conventional HMM systems," in *Proceedings of ICASSP 2000*, Vol. 3, pp. 1635–1638.

Hermansky, H., and Morgan, N. (**1994**). "RASTA processing of speech," IEEE Trans. Speech, Audio Process. **2**, 578–589.

Hermansky, H., and Sharma, S. (**1999**). "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proceedings of ICASSP 1999*, Vol. 1, pp. 289–292.

Kanedera, N., Arai, T., Hermansky, H., and Pavel, M. (**1999**). "On the relative importance of various components of the modulation spectrum for automatic speech recognition," Speech Commun. **28**, 43–55.

Kleinschmidt, M. (**2003**). "Localized spectro-temporal features for automatic speech recognition," in *Proceedings of Eurospeech 2003*, pp. 2573–2576.

Kleinschmidt, M., and Gelbart, D. (**2002**). "Improving word accuracy with Gabor feature extraction," in *Proceedings* of Interspeech 2002, pp. 25–28.

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

Mesgarani, N., David, S., Fritz, J., and Shamma, S. (**2008**). "Phoneme representation and classification in primary auditory cortex," J. Acoust. Soc. Am. **123**, 899–909.

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

Mesgarani, N., Thomas, S., and Hermansky, H. (**2010**). "A multistream multiresolution framework for phoneme recognition," in *Proceedings of Interspeech 2010*, pp. 318–321.

Meyer, B. T., Brand, T., and Kollmeier, B. (**2011b**). "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes," J. Acoust. Soc. Am. **129**, 388–403.

Meyer, B., and Kollmeier, B. (**2011a**). "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," Speech Commun. **53**, 753–767.

Nemala, S. K., and Elhilali, M. (**2010**). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. **127**, 1817.

Pearce, D., and Hirsch, H. (**2000**). "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of ICSLP 2000*, Vol. 4, pp. 29–32.

Qiu, A., Schreiner, C., and Escabi, M. (**2003**). "Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition," J. Neurophysiol. **90**, 456–476.

Schädler, M. R. (**2011**). "Gabor filter bank (GBFB) feature extraction reference implementation in MATLAB," available at http://medi.uni-oldenburg.de/GBFB (Last visited February 27, 2012).

Schwarz, R., Anastasakos, T., Kubala, F., Makhoul, J., Nguyen, L., and Zavaliagkos, G. (**1993**). "Comparative experiments on large vocabulary speech recognition," in *Proceedings of the Workshop on Human Language Technology*, pp. 75–80.

Varga, A., and Steeneken, H. J. M. (**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 Commun. **12**, 247–251.

Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., and Kollmeier, B. (**2005**). "Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines," in *Proceedings of Eurospeech/Interspeech 2005*, pp. 1273–1276.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (**2001**). *The HTK Book, version 3.1* (Cambridge University Engineering Department, Cambridge, UK), pp. 1–271.

Zhao, S., and Morgan, N. (**2008**). "Multi-stream spectro-temporal features for robust speech recognition," in *Proceedings of Interspeech 2008*, pp. 898–901.

Zhao, S., Ravuri, S., and Morgan, N. (**2009**). "Multi-stream to many-stream: Using spectro-temporal features for ASR," in *Proceedings of Interspeech 2009*, pp. 2951–2954.