Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition

Bernd T. Meyer *, Birger Kollmeier

Medical Physics, Carl-von-Ossietzky Universität Oldenburg, D-26111 Oldenburg, Germany

Abstract

The effect of bio-inspired spectro-temporal processing for automatic speech recognition (ASR) is analyzed for two different tasks with focus on the robustness of spectro-temporal Gabor features in comparison to mel-frequency cepstral coefficients (MFCCs). Experiments aiming at extrinsic factors such as additive noise and changes of the transmission channel were carried out on a digit classification task (AURORA 2) for which spectro-temporal features were found to be more robust than the MFCC baseline against a wide range of noise sources. Intrinsic variations, i.e., changes in speaking rate, speaking effort and pitch, were analyzed on a phoneme recognition task with matched training and test conditions. The sensitivity of Gabor and MFCC features against various speaking styles was found to be different in a systematic way. An analysis based on phoneme confusions for both feature types suggests that spectro-temporal and purely spectral features carry complementary information. The usefulness of the combined information was demonstrated in a system using a combination of both types of features which yields a decrease in word-error rate of 16% compared to the best single-stream recognizer and 47% compared to an MFCC baseline.

© 2010 Elsevier B.V. All rights reserved.

Keywords: Spectro-temporal feature extraction; Automatic speech recognition; Robustness; Intrinsic variability

1. Introduction

The large gap in performance between human speech recognition (HSR) and advanced automatic speech recognition (ASR) is most drastically encountered in adverse acoustic conditions and prohibits ASR technology from being widely used. Consistently, humans outperform machines by at least an order of magnitude (Lippmann, 1997). In recent studies, the gap between human and automatic recognizers was found to be somewhat smaller, but error rates are still more than 150% higher for ASR than for HSR for a simple phoneme recognition task (Meyer et al., 2007). Human listeners outperform ASR systems not only in acoustically challenging situations (e.g., in the presence of noise or competing talkers), but also when previously unknown clean speech is to be recognized. Intrinsic factors such as gender, speaking rate and style, dialect, accent, and vocal effort contribute to the vast variability the human auditory system can cope with much better than current speech recognizers. Hence, finding auditory models that adequately model speech perception has to include the difficult task of modeling human robustness against these intrinsic variations. Our attempt to narrow the gap between human and automatic speech recognition is thus motivated by the idea of transferring auditory processing principles from the human auditory system to ASR.

While many cognitive aspects of speech perception still lie in the dark, there is much progress in the research on signal processing of the auditory system. Many properties of the peripheral part of the auditory system are well understood and are already applied in current ASR recognizers, such as non-linear frequency scaling or amplitude compression. However, findings regarding the central part
of the auditory system are not considered in conventional ASR systems. For example, a number of physiological experiments in different mammal species showed that a large percentage of neurons in the primary auditory cortex (A1) respond differently to upward- versus downward-moving ripples in the spectrogram of the input. Furthermore, individual neurons are sensitive to specific spectro-temporal modulation frequencies in the incoming sound signal (Depireux et al., 2001).

The neurophysiological data fit well with psychoacoustic experiments on early auditory features: In (Kaernbach, 2000) a psychophysical reverse correlation technique was applied to masking experiments with semi-periodic white noise. The resulting basic auditory feature patterns were distributed in time and frequency and in some cases consisted of several unconnected parts, very much resembling the spectro-temporal receptive field (STRF) of cortical neurons, i.e., the model representation of the excitatory and inhibitory neurons in the auditory cortex. The STRF of a neuron is an estimate for the spectro-temporal representation of a stimulus that optimally drives that neuron. The STRFs often clearly exceed one critical band in frequency, have multiple peaks and also show tuning to temporal modulation (cf. the example in Fig. 1). Still, the STRF patterns are mainly localized in time and frequency, generally spanning at most 250 ms and one or two octaves, respectively. In the visual cortex, comparable neural tuning to spatially complex and time-varying patterns were measured with (moving) orientated grating stimuli. The results match very well with two-dimensional Gabor functions (DeValois and De-Valois, 1980), which are composed of a Gaussian envelope and a sinusoidal carrier. Gabor functions have also been used to approximate auditory STRFs as a sum of time-frequency separable Gabor functions (Qiu et al., 2003). Response patterns derived from STRFs were shown to correlate with articulatory features of phonemes (such as voicing or place or articulation) and resulted in confusion matrices similar to confusions from human listeners when used as features for ASR (Mesgarani et al., 2007).

The physiological findings have inspired a number of ASR studies that make explicit use of spectro-temporal features instead of relying on the common extraction of purely spectral features and adding delta and double-delta derivatives: Kleinschmidt et al. used Gabor functions as a simple model for STRFs to compute features for automatic speech recognizers (Kleinschmidt, 2002, 2003; Kleinschmidt and Gelbart, 2002). The parameters for the Gabor filters (such as spectral and temporal extent and modulation frequencies) were optimized in a stochastic process with physiologically motivated constraints. Features obtained from these filters improved digit recognition scores compared to an MFCC baseline by 56% on average. This approach of spectro-temporal processing by using localized sinusoids matches the neurobiological data and also incorporates other features as special cases: purely spectral Gabor functions perform an analysis similar to Mel-frequency cepstral coefficients (MFCCs) – modulo the windowing function – and purely temporal ones can resemble TRAPS or the RASTA impulse response and its derivatives (Hermansky, 1998) in terms of temporal extent and filter shape.

In a related study, a large number of Gabor features was used to cover a wider range of modulation frequencies, which were subsequently processed with multiple non-linear neural networks to merge the spectro-temporal feature streams (Zhao and Morgan, 2008). Heckmann et al. (2008) proposed hierarchical spectro-temporal features based on Gabor filtering for ASR. Similar to the work by Kleinschmidt et al. they used a statistical process to select filter parameters that yield optimal recognition performance. Gabor filters were applied to the output of a Gammatone filterbank, which resulted in localized spectro-temporal features. These features were then combined to cover a wide range of frequencies and temporal ranges. Spectro-temporal processing based on a 2D discrete cosine transform (DCT) was analyzed by Ezzat et al. (2007). The 2D-DCT was applied to patches of the short-time Fourier transform. By using only lower coefficients as basis for ASR features, relevant information including spectro-temporal patterns were extracted from speech. Despite the different approaches in these studies, spectro-temporal features were found to improve the baseline recognizers in stream-combination experiments by approximately 20–30%.

Even though the features described so far have the potential of reducing the human–machine gap in ASR discussed above for extrinsic variations (i.e., speech-in-noise), it is unclear if they show the same potential for intrinsic variations. Both properties would be a necessary prerequisite for including spectro-temporal features in advanced auditory models for human speech perception as well as ASR systems in order to close the human–machine performance gap. This study is based on the work by Kleinschmidt et al. and focuses on the robustness against the variability in spoken language. We investigated whether explicit use of spectro-temporal information increases the

![Fig. 1. Spectro-temporal receptive field (STRF) of a neuron in the primary auditory cortex of the Mongolian gerbil (adapted from Happel et al., 2008). The STRF is an estimate of the stimulus that optimally elicits a neuronal response. Dark and light areas denote excitatory and inhibitory regions, respectively, which are composed of localized, disconnected patches.](image-url)
overall robustness against extrinsic and intrinsic factors. Additionally, we report on complementary information of MFCCs and spectro-temporal features, and on the theoretical and practical improvements resulting from a combination of feature types.

2. Feature types

2.1. Spectro-temporal Gabor features

Gabor features are calculated by processing a spectro-temporal representation of the input signal with a number of 2-D modulation filters. The filtering is performed by correlation over time of each input frequency channel with the corresponding part of the Gabor function (centered on the current frame and desired frequency channel) and a subsequent summation over frequency. This yields one output value per frame per filter and is equivalent to a 2-D correlation of the input representation with the complete filter function and a subsequent selection of the desired frequency channel of the output. In this study, log mel-spectrograms serve as input features for the feature extraction. This was chosen for its widespread use in ASR and because the logarithmic compression and mel-frequency scale may be considered a very simple model of peripheral auditory processing. Even though features from a more sophisticated auditory model could have been used for spectral decomposition and temporal envelope compression (such as, e.g., the perception model used by Tchorz and Kollmeier (1999)) the usage of the standard preprocessing stage allows for a better separation of the observed effects between (temporal and spectral) preprocessing and the spectro-temporal feature extraction which is of primary interest here.

The two-dimensional complex Gabor function \( G(n, k) \) is defined as the product of a truncated Gaussian envelope \( g(n, k) \) and the complex sinusoidal function \( s(n, k) \). Alternatively, the filter can be designed as the product of a Hanning envelope \( h(n, k) \) and \( s(n, k) \), which was shown to result in improved filter characteristics and improved ASR scores (Meyer and Kollmeier, 2008). In this study we therefore use modified Gabor filters with a Hanning envelope.

The envelope width is defined by the window lengths \( W_n \) and \( W_k \), while the periodicity is defined by the radian frequencies \( \omega_n \) and \( \omega_k \) with \( n \) and \( k \) denoting the time and frequency index, respectively. The two independent parameters \( \omega_n \) and \( \omega_k \) allow the Gabor function to be tuned to particular directions of spectro-temporal modulation, including diagonal modulations. Further parameters are the centers of mass of the envelope in time and frequency \( n_0 \) and \( k_0 \). In this notation, the Hanning envelope \( h(n, k) \) is defined as

\[
h(n, k) = 0.5 - 0.5 \cos \left( \frac{2\pi(n - n_0)}{W_n + 1} \right) \cdot 0.5 \cos \left( \frac{2\pi(k - k_0)}{W_k + 1} \right)
\]

and the complex sinusoid \( s(n, k) \) as

\[
s(n, k) = \exp[i\omega_n(n - n_0) + i\omega_k(k - k_0)].
\]

The envelope width is chosen depending on the modulation frequency \( \omega_x \) (with \( x = k \) or \( x = n \)) respective the corresponding period \( T_x \), either with a fixed ratio \( v_x = T_x/2W_x = 1 \) to obtain a 2D wavelet prototype or by allowing a certain range \( v_x = 1, \ldots, 3 \) with individual values for \( T_x \) being optimized in the automatic feature selection process. For time dependent features, \( n_0 \) is set to the current frame, leaving \( k_0, \omega_k \) and \( \omega_n \) as free parameters. From the complex results of the filter operation, real-valued features are obtained by using the real, imaginary or absolute part only. In this case, \( W_x \) replaces \( \omega_x = 0 \) as a free parameter, denoting the extent of the filter, perpendicular to its direction of modulation.

2.1.1. Feature set optimization

In order to apply Gabor filters to the problem of speech recognition, parameter sets from a large number of possible combinations need to be determined. Feature set optimization was carried out by a modified version of a Feature-finding Neural Network (FFNN). It consists of a linear single-layer perceptron in conjunction with an optimization rule for the feature set (Gramss and Strube, 1990). The linear classifier guarantees fast training, which is necessary because in this wrapper method for feature selection, the importance of each feature is evaluated by the increase of RMS classification error after its removal from the set. This ‘substitution rule’ method (Gramss, 1991) requires iterative re-training of the classifier and replacing the least relevant feature in the set with a randomly drawn new one. When the filter set is optimized with a database containing isolated words without phoneme labels, a temporal integration of features is carried out by simple summation of the feature vectors over the whole utterance. This results in one feature vector per utterance as required for the linear net. The FFNN approach has been successfully applied to digit recognition in combination with Gabor features in the past (Kleinschmidt, 2002; Kleinschmidt and Gelbart, 2002).

The Gabor filter set used in this study was obtained by using the FFNN with the ZIFKOM German digit data,

---

1 This issue may be solved implicitly by automatic learning in neural networks with a spectrogram input and long time windows of e.g. 1s. However, this is computationally expensive and prone to overfitting, as it requires large amounts of training data, which are often unavailable. By putting further constraints on the spectro-temporal patterns, the number of free parameters can be decreased by several orders of magnitude. This is the case when a specific analytical function, such as the Gabor function, is explicitly demanded. This approach narrows the search to a certain sub-set and thereby some important features might be ignored. However, neurophysiological and psychoacoustical knowledge can be exploited for the choice of the prototype, as it is done here.

---

which contains single digit utterances spoken by 100 female and 100 male speakers. For distortion, three types of noise were added to the utterances (a) un-modulated speech-shaped noise (CCITT G.227), with a spectrum similar to the long-term spectrum of speech, (b) real babble noise recorded in a cafeteria situation and (c) speech-like shaped and modulated noise (ICRA noise signal 7 (Dreschler et al., 2001)). Before mixing, speech and noise signals were band-pass filtered to 300–4000 Hz, roughly corresponding to the telephone band. The SNRs were chosen analogous to the AURORA 2 framework, i.e., SNRs from 20 to −5 dB were employed. Temporal and spectral modulation frequencies for the filters were randomly chosen in an interval from 2 to 50 Hz and 0.06–0.5 cycles/octave, respectively. Boundary conditions for the spectral extent of the filter guaranteed that even at low modulation frequencies the filters did not exceed 23 frequency channels or 101 time frames (corresponding to 1 s filter length). Five hundred iterations were used for the selection process and approximately 30 min of speech were processed by the FFNN in real-time on a typical workstation computer. The optimized filter set contains 80 filter functions; the 15 filters which contributed most to the classification performance of the speech data are shown in Fig. 3, and the according parameters are presented in Table 1. The properties of the complete filter set are depicted in Fig. 4, which includes the distribution of temporal and spectral modulation frequencies. The filter set employed in this study and a front-end to calculate Gabor features from speech signals can be obtained from http://www.icsi.berkeley.edu/speech/papers/gabor/.

2.1.2. Non-linear transformation

The original 80-dimensional filter output was processed by a Tandem system (Hermansky et al., 2000) as shown in Fig. 5: In a first step, the feature vectors were online normalized and combined with delta and double-delta derivatives before using them as input to a non-linear neural net (or multi-layer perceptron (MLP)). The MLP was provided by the QuickNet software package (http://www.icsi.berkeley.edu) and had 3 · 80, 1000 and 56 neurons in input, hidden and output layer, respectively. It was trained on the TIMIT phone-labeled database with artificially added noise. The resulting posteriors were decorrelated using a principal component analysis (PCA) which yields 56-dimensional feature vectors. These were used as input features to a Hidden Markov model (see Section 3.2).

2.2. MFCC features

Mel-frequency cepstral coefficients (MFCCs) have been chosen as baseline for this series of experiments. For the computation of MFCCs (Davis and Mermelstein, 1980), a pre-emphasis is applied to the signal before calculating the smoothed short-time Fourier transform (STFT). Each frame is then processed by a mel-filterbank (which approximates the response of the human ear), compressed with the logarithm and transformed to cepstral parameters using an inverse discrete cosine transformation. By selecting the lower cepstral coefficients, only the coarse spectral structure is retained. This processing results in mostly decorrelated features.

For the presented experiments, MFCC features were calculated using the rastamat Matlab toolbox (Ellis, 2003) with parameters that resemble feature extraction from the HTK software (Young et al., 1995), i.e., the filter bank used 20 frequency channels; the 13-dimensional features were concatenated with delta and acceleration
coefficients. Signals with 16 kHz bandwidth were used as input to the front-ends.

3. Methods

3.1. Speech databases, training and testing sets

The robustness of ASR features against extrinsic variations, i.e., additive noise or channel distortions is often assessed by the performance loss produced by mismatches between training and testing noise types or SNRs. Similarly, the robustness against intrinsic variations may be evaluated by training the recognizer on normally spoken utterances and testing it on utterances that cover a wider range of variabilities. In this work, two databases were used for ASR experiments to cover both aspects of variability.

3.1.1. AURORA 2 database

The AURORA 2 framework was used to assess the impact of additive noise sources. The database contains strings of connected digits from the TIDigits database (Leonard, 1984) to which various noise types were added at SNRs ranging from −5 dB to 20 dB in 5 dB-steps. The framework provides two training modes: ‘Multi-condition’ refers to training the recognizer with clean and noisy signals, where four noise types are used (suburban train, crowd of people (babble), car, and exhibition hall). For ‘clean condition’ training, only utterances without additional noise have been employed.

The test set covers eight noise types at SNRs from −5 dB to 20 dB, as well as clean speech. Four of these noises are the same as for multi-condition training, while the remaining noises (restaurant, street, airport and train station) are not used during training. Therefore, the effect of matched vs. mismatched training and testing can be investigated. The test set also includes speech signals filtered with a telephone bandpass characteristic before applying the noises suburban train and street, taking channel transmission effects into account.

Table 1

Parameters of the 15 most important Gabor filters depicted in Fig. 1. The table specifies the center frequency $f_c$, temporal and spectral modulation frequencies $\omega_m$ and $\omega_k$, and the filter type (either purely temporal, purely spectral, or spectro-temporal (upward or downward)). The mode defines which part of the complex filter output was used (either real or imaginary part or the absolute value).

<table>
<thead>
<tr>
<th>$f_c$</th>
<th>$\omega_m$</th>
<th>$\omega_k$</th>
<th>Type</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>0.039</td>
<td>0</td>
<td>Temporal</td>
<td>Mag</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.124</td>
<td>Spectral</td>
<td>Real</td>
</tr>
<tr>
<td>8</td>
<td>−0.04</td>
<td>0.069</td>
<td>STup</td>
<td>Real</td>
</tr>
<tr>
<td>2</td>
<td>0.021</td>
<td>0</td>
<td>Temporal</td>
<td>Real</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0.081</td>
<td>Spectral</td>
<td>Imag</td>
</tr>
<tr>
<td>23</td>
<td>0.049</td>
<td>0</td>
<td>Temporal</td>
<td>Imag</td>
</tr>
<tr>
<td>12</td>
<td>0.023</td>
<td>0</td>
<td>Temporal</td>
<td>Imag</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0.216</td>
<td>Spectral</td>
<td>Mag</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0.135</td>
<td>Spectral</td>
<td>Real</td>
</tr>
<tr>
<td>2</td>
<td>0.03</td>
<td>0</td>
<td>Temporal</td>
<td>Real</td>
</tr>
<tr>
<td>18</td>
<td>0.048</td>
<td>0</td>
<td>Temporal</td>
<td>Mag</td>
</tr>
<tr>
<td>6</td>
<td>−0.022</td>
<td>0.123</td>
<td>STup</td>
<td>Mag</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0.077</td>
<td>Spectral</td>
<td>Imag</td>
</tr>
<tr>
<td>9</td>
<td>0.102</td>
<td>0.071</td>
<td>STDown</td>
<td>Real</td>
</tr>
<tr>
<td>14</td>
<td>−0.045</td>
<td>0.125</td>
<td>STup</td>
<td>Imag</td>
</tr>
</tbody>
</table>

Fig. 3. Gabor filter functions obtained with the Feature Finding Neural Network. The figure shows the 15 most relevant filters from a set with a total of 80 filters.

In order to evaluate the robustness of a system, results for the clean trained HMM are of special interest, as the HMM models do not contain any specific information about possible distortions in this case. Therefore, the scores obtained with this training mode are a good measure for the invariance of features against the noise types in the test set.

3.1.2. Oldenburg Logatome Corpus

The Oldenburg Logatome Corpus (OLLO) is a database that was recorded for speech intelligibility tests with human listeners and for experiments with automatic classifiers (Wesker et al., 2005). It consists of non-sense utterances or logatomes, i.e., words without semantic meaning which comply with phonetic and phonotactic rules. The logatomes are composed of triplets of vowels (V) and consonants (C), with the outer phonemes being identical. Fifty speakers recorded 70 VCVs and 80 CVCs with different speaking styles, efforts and speaking rates, thus enabling an analysis of the effect of such intrinsic variations of speech. During the recordings, participants were asked to speak each logatome either normally or in one of five variations. The properties of the database are listed in Table 2. For details on the OLLO corpus, the reader is referred to (Wesker et al., 2005).

For the ASR experiments in this study, subsets of the OLLO database were selected for training and testing the recognizer. The variabilities include fast and slow speaking rate, high and low speaking effort (conditions ‘loud’ and ‘soft’), and condition ‘question’ which refers to utterances with rising pitch. Additionally, normally spoken utterances were included as reference condition. The variabilities were
equally distributed in the test set. The ASR task was to recog- 
nize one of 14 middle consonants or one of 10 central 
vowels. The use of a phoneme recognition task allows for 
an analysis of the errors on phoneme level (e.g., consonant 
vs. vowel confusions) to gain some insight into which prop- 
erties of speech sounds result in correct and incorrect 
classification.

Utterances of the OLLO database from three male and 
three female German speakers (without regional dialect) 
served as training data, logatomes from the four remaining 
speakers without dialect were used for the testing proce- 
dure. The chosen segmentation of the corpus results in a 
speaker- and gender-independent ASR system. While the 
training set contained only normally spoken logatomes, 
the test set additionally contained utterances with the 
aforementioned variations (conditions ‘fast’, ‘slow’, ‘loud’, 
‘soft’, ‘question’).

The training and testing was carried out with noisy sig- 
nals for which a speech-shaped stationary noise (Dreschler 
et al., 1999) was added to the utterances at SNRs ranging 
from −10 to 10 dB in 5 dB-steps. Since the focus of this 
experiment was on intrinsic variations, the same SNR 
was chosen for training and testing. The SNR was calcu- 
lated by relating the root-mean-square (rms) value of the 
speech segments of each audio signal and the rms value 
of the masking noise of equal length. A simple voice detec- 
tion algorithm based on an energy criterion was used to 
exttract connected speech segments. Additionally, the classi-
 fier was trained and tested with clean speech.

3.2. Automatic recognizers

3.2.1. HTK baseline recognition system

Gabor and MFCC features were used to train and test 
recognition systems based on Hidden Markov models 
(HMM), implemented in the HMM toolkit (Young et al., 
1995). For experiments with the AURORA 2 database, 
the classifier was configured according to (Hirsch and 
Pearce, 2000), i.e., the HMM used 16 states per word and 
three Gaussian mixtures per state, which are connected 
by left-right-models that do not allow for skipping states. 
For experiments with the Oldenburg Logatome Corpus, 
the task was defined as recognition of the central phoneme 
in the CVCs and VCVs, mimicking earlier experiments with 
human listeners based on the OLLO corpus (Meyer and 
Wesker, 2006). Logatomes with the same outer phoneme 
were used to train and test single HMMs (based on 
HTK) which were subsequently used to classify the central 
phoneme in CVCs and VCVs, i.e., confusion occurred only 
for central phonemes. Note that in this test setup, confu-
sions between the consonant and the vowel group cannot 
occur. The HTK was configured with three states per pho-
neme and eight mixtures per state.

3.2.2. Philips Continuous ASR system

Additionally, the performance of Gabor features vs. 
MFCCs was tested using a more advanced recognition sys-
tem, i.e., the Philips Continuous ASR system (Lieb and 
Fischer, 2002). This classifier was chosen for two reasons: 
First, it was investigated whether Gabor features can 
increase performance in a recognition system that incorpo-
rates denoising techniques as well as methods to improve 
auditory modeling such as discriminative training for 
HMMs. Second, the Philips ASR system provides methods 
to combine feature streams that were used for an analysis 
regarding feature complementarity of purely spectral and 
spectro-temporal features, as described in Section 4.4.

The standard feature extraction stage for the Philips 
ASR system is based on MFCCs (12 cepstral coefficients 
with delta features which yields 24-dim. feature vectors), 
which are combined with an HMM classifier. It features 
read extraction techniques such as non-linear spectral subtrac-
tion, noise-masking or linear discriminant analysis (LDA) 
and classification based on discriminative training. Gabor 
and MFCC features were used as input to the recognizer 
both individually and with a combination of feature 
streams. As for the experiments based on HTK, the sets 
for training and testing as proposed in the AURORA 2 
framework were used.

4. Results

This section presents the outcome of ASR experiments 
with MFCC features and the optimized Gabor filter set 
described in Section 2.1.1. The section is structured as fol-
ows: First, the results for the experiments aiming at robustness against extrinsic and intrinsic variations are 
reported in Sections 4.1 and 4.2, respectively, and the 
results are compared in Section 4.3. Both experiments were 
carried out with the standard HTK recognition system (cf.
Section 3.2.1). The results regarding the complementarity of spectral and spectro-temporal features (performed with the Philips recognition system) are presented in Section 4.4.

4.1. Effect of extrinsic variations

The effect of extrinsic variations was investigated using the AURORA 2 task with various additive noise conditions. The HTK reference recognizer was trained and tested with 39-dimensional MFCC features (as proposed in (Hirsch and Pearce, 2000)), as well as with spectro-temporal features. Gabor features improved the baseline for all noise types as shown in Table 3. The presented scores were obtained by averaging over the SNRs from 20 dB to 0 dB. Additionally, relative reductions in WER (as commonly reported for the AURORA 2 task) are presented. On average, errors were reduced by 11.5% relative for multi-condition training. When the recognizer is trained with clean utterances and tested with a variety of noise types the difference between the feature types becomes more noticeable, as the baseline error rate is reduced by 57.8% relative when Gabor features are used.

Gabor consistently outperform MFCCs for noise types which are used in multi-condition training (first four noise types in Table 3), as well as for noises not being used for the training (noises 5–8), and noisy signals which have been filtered with realistic frequency characteristics (‘Subway M’ and ‘Street M’). The average benefit is consistent over these conditions, i.e., Gabor features have a similar sensitivity against mismatches of the employed training noises as MFCCs. This result confirms findings from other studies investigating spectro-temporal features, i.e., the noise robustness of recognizers based on spectral processing of speech can be improved by including spectro-temporal information.

4.2. Effect of intrinsic variations

Intrinsic variability and its impact on ASR phoneme recognition were analyzed with the HTK system, which was trained and tested with utterances from the Oldenburg Logatome Corpus (cf. Section 3.2.1). Phoneme recognition rates depending on speech intrinsic variations are shown in Table 4 for MFCC and Gabor features. Gabor features that were used as direct input to the HMM (i.e., the non-linear transformation with the neural net as described in Section 2.1.2 was omitted) produced scores between MFCC and non-linearly transformed Gabor features and are not shown in the table.

The overall performance with matched training and testing is similar for cepstral and Gabor features: When averaging over all SNRs and variabilities, the phoneme accuracy is 51.0% (MFCCs) and 53.0% (Gabor). Intrinsic variations degrade ASR performance compared to the reference condition for both feature types: The performance drop averaged over all SNRs is 13.6% and 14.9% absolute for MFCCs and Gabor features, respectively. The relative increase in terms of word-error rate is 36.2% (MFCCs) and 42.9% (Gabor). Rather large differences between the ASR feature types are observed for the conditions ‘loud’ and ‘question’: When clean utterances are used for training and testing, the relative MFCC error rates are at least 25% higher than those of Gabors. On average, Gabor features were found to be less sensitive to changes in speaking effort (‘loud’ and ‘soft’) and style (‘question’) than MFCCs. On

<table>
<thead>
<tr>
<th>Multi condition training</th>
<th>Restaurant</th>
<th>Street</th>
<th>Airport</th>
<th>Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subway Babble Car Exhibition</td>
<td>86.6</td>
<td>87.8</td>
<td>88.3</td>
<td>86.2</td>
</tr>
<tr>
<td>Subway M</td>
<td>83.5</td>
<td>85.7</td>
<td>87.0</td>
<td></td>
</tr>
<tr>
<td>Street M</td>
<td>88.1</td>
<td>87.3</td>
<td>88.6</td>
<td></td>
</tr>
<tr>
<td>MFC</td>
<td>4.8</td>
<td>8.0</td>
<td>10.8</td>
<td>11.1</td>
</tr>
<tr>
<td>Gabor</td>
<td>28.1</td>
<td>11.1</td>
<td>11.5</td>
<td></td>
</tr>
<tr>
<td>Rel. reduction WER</td>
<td>7.3</td>
<td>5.9</td>
<td>16.5</td>
<td>11.4</td>
</tr>
<tr>
<td>MP MCC</td>
<td>8.1</td>
<td>4.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gabor</td>
<td>8.3</td>
<td>7.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rel. reduction WER</td>
<td>10.5</td>
<td>8.6</td>
<td>20.2</td>
<td></td>
</tr>
<tr>
<td>Clear condition training</td>
<td>Subway Babble Car Exhibition</td>
<td>Restaurant</td>
<td>Street</td>
<td>Airport</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------------</td>
<td>------------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>Subway Babble Car Exhibition</td>
<td>50.1</td>
<td>60.7</td>
<td>49.6</td>
<td>53.1</td>
</tr>
<tr>
<td>Subway M</td>
<td>65.3</td>
<td>66.7</td>
<td>58.1</td>
<td></td>
</tr>
<tr>
<td>Street M</td>
<td>85.5</td>
<td>82.7</td>
<td>82.4</td>
<td></td>
</tr>
<tr>
<td>MFC</td>
<td>7.5</td>
<td>83.4</td>
<td>80.4</td>
<td>82.5</td>
</tr>
<tr>
<td>Gabor</td>
<td>58.3</td>
<td>48.1</td>
<td>57.8</td>
<td></td>
</tr>
<tr>
<td>Rel. reduction WER</td>
<td>58.3</td>
<td>57.7</td>
<td>61.1</td>
<td>62.7</td>
</tr>
<tr>
<td>MP MCC</td>
<td>58.7</td>
<td>53.4</td>
<td>66.0</td>
<td></td>
</tr>
<tr>
<td>Gabor</td>
<td>83.5</td>
<td>80.4</td>
<td>84.1</td>
<td></td>
</tr>
<tr>
<td>Rel. reduction WER</td>
<td>59.8</td>
<td>58.1</td>
<td>53.2</td>
<td></td>
</tr>
</tbody>
</table>
the other hand, spectro-temporal features were more affected by variations of speaking rate (‘fast’ and ‘slow’).

The features’ performance with respect to consonant and vowel recognition was also analyzed. The overall performance (depending on intrinsic factors) was therefore broken down into consonant and vowel recognition rates (Fig. 6). Results for a high- and a low-noise condition (−5 dB and +10 dB SNR) are presented. The figure shows that – in the presence of intrinsic variations – MFCCs performed better (4.7% absolute on average) with respect to consonant recognition, the only exceptions being the classification for the categories ‘loud’, ‘soft’, and ‘question’ at the high SNR. On the other hand, the vowel recognition rates were higher (11.6% absolute, averaged over both SNRs) with spectro-temporal Gabor features. This trend was found throughout all intrinsic variations. This suggests that both feature types carry complementary information with respect to consonant/vowel recognition, which is discussed in Section 5.

4.3. Comparison of extrinsic and intrinsic factors

The performance degradation caused by intrinsic variations can be compared to the degradation caused by additive noise. Fig. 7 shows the relative increase of phoneme error rates for both feature types over the SNR. The
The relative increase in each panel is related to normally spoken utterances without noise. When the recognizer is trained and tested with clean speech, changes of speaking style and effort result in an average increase of 55.5% and 50.0% for MFCC and Gabor features, respectively. The same overall degradation is caused by noise being added at approximately 0 dB for both features types.

Although the average increase of error is similar for both feature types, the variance of differences caused by intrinsic variations are larger for MFCCs than for Gabor features: For example, changes in speaking rate have a relatively small effect on MFCCs with a relative increase in error of 13% for condition ‘slow’. On the other hand, loud speaking style results in an increase of errors by a factor of two. With degradation between 37% (‘question’) and 61% (‘fast’), the fluctuations observed for Gabor features are much smaller.

The influence of intrinsic and extrinsic factors was also examined at a ‘microscopic’ level: The overall scores for noisy speech at 0 dB SNR and for clean speech with all intrinsic variations from the OLLO corpus were found to be similar. Hence, we compared the phoneme confusions of these two conditions. The addition of stationary noise is excepted to severely degrade consonant performance due to energetic masking (Barker and Cooke, 2007), while vowels are presumably less affected by the masker. The recognition rates for both conditions are shown in Fig. 8. The comparison shows that the errors due to extrinsic and intrinsic factors show very large differences: The degradation of consonants is notable for both feature types in the noisy condition, but not in the presence of intrinsic variations. The correlations between the conditions were not significant (MFCCs: correlation coefficient: \( r = 0.22 \), p-value: \( p = 0.28 \). Gabor features: \( r = -0.38 \), \( p = 0.06 \)).

Fig. 7. Relative increase of phoneme error rates resulting from higher masking level and speech intrinsic variations for two ASR feature types, depending on the SNR used for training and testing. The horizontal dashed line denotes the average increase of error \( E_V \) induced by intrinsic variability for clean utterances. The vertical line highlights the SNR at which the error of normally spoken utterances reaches \( E_V \).

Fig. 8. Comparison of phoneme recognition rates in the presence of extrinsic and intrinsic factors. Experiments with similar overall scores are compared for MFCC and Gabor features (left and right panel, respectively).

The different error patterns suggest that different techniques are required to cope with the variability introduced by intrinsic and extrinsic factors.

4.4. Complementarity of spectral and spectro-temporal features

When using delta coefficients as additional features, the MFCC feature vector captures some information about the temporal dynamics of the speech signal; however, truly spectro-temporal information is not included on feature level. While this limitation is a theoretical disadvantage of MFCCs, it is often difficult to achieve improvements for tuned ASR systems with completely new features. We therefore investigated if both feature types carry complementary information and if a combination of Gabor and MFCC features is a promising approach. For these experiments, we chose results obtained with the Philips continuous recognizer (Lieb and Fischer, 2002) with an improved feature extraction stage that uses noise suppression before feature calculation (cf. Section 3.2.2). The recognition task was the same as for the HTK recognizer, i.e., digit classification within the AURORA 2 framework, while the baseline feature extraction was slightly modified (i.e., 12-dimensional cepstral features with delta features were used).

The intersection of misclassified digit tokens $E$ from both systems was chosen as a measure for complementary information: $I_{err} = E_{Gabor} \cap E_{MFCC}$. The smaller $I_{err}$ is, the smaller is the error rate of an (imaginary) perfect classifier that can use the MFCC or the Gabor feature information, and thus only produces an error if a digit was misclassified by both single-stream systems. A low error rate of such a perfect or ‘oracle’ system represents a high complementarity of feature streams. Insertions and deletions are included in $I_{err}$ if an insertion or deletion occurs at the same position of the transcribed string of digits. The word accuracies of both feature types and the oracle system are shown in Table 5. Performance obtained with Gabor was between scores for denoised and original MFCC features. When denoised MFCCs are used as baseline, the perfect knowledge scenario decreases the error rates about 55% relative. Note that the scores slightly differ from the results presented in Table 3 since a different back-end was employed.

These results motivated a combination of feature streams: Denoised MFCCs were concatenated with Gabor features and used to train and test the Philips recognizer, as described in Section 3.2. The Philips recognizer as described in (Lieb and Fischer, 2002) and in Section 3.2.2 was optimized on 12-dimensional cepstral coefficients with additional delta features. For reasons of comparability with results from (Lieb and Fischer, 2002), the tests investigating the complementarity were carried out with 24-dimensional features. Therefore, Gabor features were reduced to 24 dimensions by selecting the first 24 components of the PCA-transformed feature vector. The concatenation with MFCCs resulted in 48-dimensional vectors, which were transformed to 24-dimensional vectors using a linear discriminant analysis. It is possible that the limitation to 24 Gabor feature components results in a suboptimal performance with this feature type and a higher dimensionality would improve the scores. This was however not tested, since in this set of experiments, the aim was to compare classifiers with an identical number of model parameters.

The result of the stream-combination experiment is shown in Table 5 (row c): The word accuracy was improved for clean and multi-condition training when using concatenated Gabor features and denoised MFCCs. The scores correspond to relative reductions of 16% in average and 23% for multi-condition training compared to denoised MFCCs. Compared to MFCCs without denoising, the average reduction in WER is over 70%. This outcome suggests that MFCCs and Gabor features carry complementary information, and spectro-temporal processing can be used to improve an advanced classification system such as the employed Philips recognition system.

5. Discussion

5.1. Robustness of Gabor features to extrinsic variations

The comparison of spectro-temporal and MFCC features showed that Gabor features obtained with the filter set used in this study provide increased robustness against a wide range of noise sources. Improvements over the AURORA 2 baseline were found when the recognizer was trained on noisy utterances. The average reduction in relative word error was 35% for the HTK recognizer and 47% for the Philips Continuous ASR system compared to MFCC features.

The increased performance might be a result of the FFNN algorithm: Since the optimization of filter parameters is carried out on speech material, the (optimized) Gabor filters can be considered as matched filters for distinctive speech features. These filters appear as relatively robust against additive noise. In contrast to this, the
spectral content of speech is well encoded by MFCC features that deliver good performance for tests with clean speech. However, this representation seems to be severely degraded in the presence of noise. It might be the spectro-temporal cues (which might be redundant for clean speech, but become more important at low SNRs), which cause the increased robustness against extrinsic variations found in these experiments.

Interestingly, the best performance with Gabor features was obtained with the filter set optimized on German digit data, although one of the recognition tasks was the classification of English utterances. Furthermore, Gabor phoneme classification performance was on par with MFCCs, even though the set of phonemes used in filter selection (ZIFKOM database) differed from the phonemes used in ASR (OLLO database): The phoneme /x/ is contained in the German word ‘acht’ (eight), but not in the logatomes from the OLLO corpus. On the other hand, several phonemes in OLLO are not contained in German digits (/g, l, m, n, p, s/). This indicates that the selection procedure is relatively insensitive towards mismatches between the phoneme inventory in feature selection and in the actual classification task, and that the Gabor filter set used in this study might be suitable for a larger group of recognition tasks without the need to optimize a new filter set for each test condition. A further parameter that might be important in this context is the duration of utterances to be recognized: While CVC and VCV utterances are relatively short (with durations of only a few hundred milliseconds for short CVCs), the duration of digit strings in AURORA 2 is considerably higher. A higher duration might be favorable for the Gabor filters that exhibit temporal integration windows of up to 1 s. Hence, it is possible that optimizing the highest filter duration considered during filter selection can increase the phoneme classification on very short utterances.

In several physiological studies (Depireux et al., 2001; Qiu et al., 2003) it was reported that a large proportion of spectro-temporal receptive fields in the auditory cortex are separable, i.e., the STRFs can be expressed as the product of a spectral and a temporal function and do therefore not contain any ‘true’ spectro-temporal patterns. However, in our experiments with separable Gabor functions, we have so far not been able to improve the results compared to the non-separable Gabor functions that have been used in this study. This indicates that at least parts of the speech information that the (optimized) Gabor features from our set are matched to cannot be represented by a simple combination of spectral and temporal properties. Instead, a time-varying spectral content (resembling, for example, a transition in fundamental frequency or formant frequency) is specifically captured by some of the Gabor features employed here.

The noise types used during the filter selection process influence the parameters of the selected filters, i.e., features may adapt to the type of noise due to the optimization process that includes a simple speech recognition task. For the filter optimization presented in this work, one stationary and two modulated noises were chosen. These were not identical to the AURORA 2 noises, but the babble noises applied for filter selection and for the recognition task had similar characteristics. However, this was not reflected in the AURORA 2 results, since the condition ‘babble’ produced an intermediate decrease of the relative error rates (Table 3). It therefore seems that the presented approach does not result in filters which are highly adapted to a specific class of noise types, but which are useful in a wider range of masking noises.

5.2. Effect of intrinsic variations

Intrinsic variations (speaking rate, style and effort) had a strong impact on ASR performance, with an overall degradation of approximately 50% for phoneme recognition in clean speech (Table 4). Our analysis showed that Gabor features calculated with the filter set specified in Section 2.1.1 differ from MFCC features regarding sensitivity against intrinsic parameters: MFCC features were less sensitive against changes in speaking rate, while the overall sensitivity against speaking effort and style was improved with Gabor features. Thus, the usage of spectro-temporal features employed in this study is not only beneficial for overall performance, but also results in different sensitivity against intrinsic variations, which could be utilized to increase robustness by combining properties of cepstral and Gabor features.

The reason for the higher error rates for fast spoken utterances might be that the optimization of the filter set was carried out on words that were spoken at normal speaking rate. Higher spectro-temporal modulation frequencies, which could be better suited to detect, e.g., formant transitions of speech at high speaking rate, may therefore not be included in the filter set. For purely spectral features, the adaptation to different rates of speech is performed in the back-end stage, whereas some of the timing information is included in the spectro-temporal features which are “frozen” in a certain word production speed used during learning. An adaptation of Gabor filters to other speaking rates could be performed by including utterances with fast and slow speaking rate (e.g., from the OLLO database) during filter selection. One of the problems to overcome in this context is the relatively low number of phonemes in the OLLO corpus, which could result in adapted filter types that might not be suitable for a broader class of recognition problems. This problem might be solved by the use of several training speech databases during filter selection, which could yield a similar noise robustness as the filter set employed in this study, and additionally may extract spectro-temporal modulations which are associated with fast and slow speaking rates.

The increase in WER due to intrinsic variations was compared to the increase due to additive noise (Fig. 8). The overall effect of the analyzed speaking style, effort and rate was approximately the same as for a stationary
masker added at 0 dB SNR to clean signals. This result was obtained in a phoneme recognition task with matched ASR training and testing. It remains to be seen if this result holds for other tasks as well (e.g., recognition of words or conversational speech) that require other speech databases with labels of such variations of speech. The comparison of phoneme confusions with intrinsic and extrinsic variabilities (Fig. 8) showed that the type of error patterns differs for the described conditions: additive noise resulted in higher consonant errors, presumably because consonants have less energy than vowels and are therefore masked by the stationary noise. Intrinsic variations had a stronger impact on vowel recognition, which may be caused by the stronger variations in the articulation of vowels due to an altered speaking style (such as, e.g., fast, slow, question,...) as opposed to the consonants that are roughly articulated in the same way.

It is an open question which cues are the most relevant for the observed differences between MFCC and Gabor features for loud speaking style. Obviously, consonant recognition seems to be more affected than vowel recognition (Fig. 6), but the perceptual cues associated with these observations are not easily accessible. Although the recordings of the OLLO database have been carried out in a quiet environment, the properties of loudly spoken utterances may be similar to those of Lombard-speech, i.e., speech that has been recorded in noisy surroundings, which results in a slight decrease of consonant duration, changes of spectral properties of fricatives, and maximum burst energy of plosives (Junqua, 1993). Changes in consonant duration do not seem to be a likely candidate, since varying speaking rate resulted in consonant degradation for Gabor features (Fig. 6). However, from the right panel in Fig. 8 it can be seen that Gabor fricative scores are above average in the presence of intrinsic variations, i.e., the spectral changes of this phoneme group might explain these observations.

Logatomes spoken with a rising pitch exhibit stronger diagonal patterns of spectral fine structure compared to the other speaking styles. This might be the reason for the improved performance of Gabor features for the condition ‘question’, since the features capture such transients even at lower SNRs if the spectro-temporal modulation frequencies coincide with the transients of the speech signal.

5.3. Interaction between the SNR and phoneme recognition

The analysis of consonant and vowel recognition scores showed that – for the majority of variabilities – MFCCs produced improved consonant scores, while Gabor features obtained with the filter set described in Section 2.1.1 produced higher vowel scores (Fig. 6). This was found for both high and low SNR conditions. The good performance of MFCCs for consonants might be due to the presence of characteristic energy clusters that are separated in the frequency domain (e.g., high-frequency elements due to frication in combination with a low-frequency stop or formant maximum characteristic for manner and place of articulation) which can be favorably detected based on the spectral envelope. On the other hand, Gabor features encode the temporal modulation and local transition of speech energy typical for formant transients for vowels, which may result in a locally increased SNR and therefore enhanced performance. The increase of the local SNR may also be the reason for the relatively high scores with Gabor features for very low SNRs (Table 4), which reflect the general robustness of this feature type against additive noise. The fact that Gabor produced slightly better consonant scores for some of the intrinsic variations (low-noise condition) might be due to the better representation of spectral fine structure and phoneme transitions (compared to cepstral coefficients). The use of such fine structure for ASR has been proposed earlier (Dimitriadis et al., 2005; Scharenborg, 2007).

5.4. Complementarity of MFCC and Gabor features

The specific Gabor filter set presented in Section 2.1.1 was not designed with a combination with MFCCs in mind, but still resulted in an increase of performance when feature streams were concatenated (Table 5). Relative WERs were reduced both in a theoretical approach (55%) as well as in a real-world scenario (16%), which demonstrated the potential of this class of physiologically motivated features. However, this is not even halfway to the WER reduction observed for the oracle system, which motivates more advanced feature combination techniques, such as LDA or the combination of multiple neural nets. These findings are in line with results from related studies that analyze spectro-temporal features: When combining high-dimensional Gabor features with MFCCs in a multi-stream environment, an improvement in WER of 30% was found (Zhao and Morgan, 2008). Similarly, a combination of RASTA-PLP (Hermansky and Morgan, 1994) and spectro-temporal features was reported to lower ASR error rates by roughly 20% (Heckmann et al., 2008).

Furthermore, the result suggests to include MFCCs in the feature selection process: The Gabor filter set with best performance exhibits 30% purely spectral and temporal filters, respectively, while 40% of the automatically defined filters are spectro-temporal. An inclusion of spectral features in the parameter definition process would presumably result in a shift away from spectro-temporal and purely temporal filters, thus increasing complementary information. Other candidates for features to be included during filter optimization are TRAPS features which account for the temporal dynamics of spoken language (Hermansky and Sharma, 1999). Note that the results obtained in this study are valid for the Gabor filter set presented in Section 2.1.1. It can be assumed that other filter sets (especially when other constraints are applied during filter selection) result in different performance and complementarity to MFCCs. The fact that cepstral features result in an improvement shows that the filter selection process performed with the FFNN does not function perfectly. Vice
versa, a close-to-perfect feature selection process should be able to combine the both the strengths of MFCCs and the Gabor features employed here.

Although the experiments based on the Oldenburg Logatome Corpus aimed at the effect of intrinsic variations, the analysis regarding consonant and vowel confusions indicated that cepstral features are better suited for consonant recognition, while spectro-temporal features improved the vowel scores. This result helps to understand why Gabor features are beneficial in a stream-combination experiment: Since the task defined in the AURORA 2 framework is to recognize digits in a wide range of SNRs and in clean speech, the different properties of feature types of spectral and spectro-temporal features increases the overall robustness. We therefore argue that Gabor features and MFCCs carry complementary information on different levels, which could be exploited in feature-stream experiments.

6. Conclusions

The most important findings from this study can be summarized as follows:

- Spectro-temporal Gabor features were found to be more robust than an MFCC-based classifier against a wide variety of extrinsic sources of variability. Small improvements were achieved when the classifier was trained and tested with a mixture of clean and noisy signals. When the training was performed with clean utterances, the reduction in word-error rate was over 50%. They are in line with other studies that analyze spectro-temporal filters for ASR, and confirm that spectro-temporal information can help to increase robustness against noise.

- The presence of intrinsic variations such as speaking rate, style and effort severely degrades the performance of ASR. In acoustically optimal conditions, the average increase of errors was over 50% for a phoneme recognition task. Purely spectral and spectro-temporal features were affected differently by these variabilities: While MFCCs were less susceptible to changes in speaking rate, the usage of spectro-temporal input for ASR resulted in performance above baseline for high and low speaking effort, as well as for utterances with rising pitch. This finding suggests a combination of spectral and spectro-temporal features to be used in future experiments on the sensitivity against intrinsic variations.

- The degradation due to intrinsic variations had a similar effect on overall phoneme recognition as a stationary, speech-shaped noise at approximately 0 dB SNR. This result was found both for cepstral and spectro-temporal features. However, the confusions of individual phonemes clearly differed, thus demonstrating that intrinsic and extrinsic variations result in different microscopic confusions.

- The errors that occur with spectro-temporal features calculated with the specific filter set used in this study are genuinely different from MFCC features. An analysis regarding complementarity showed that (a) different errors occur with each feature type on a digit recognition task and (b) these features seem to carry complementary information which might be beneficial to consonant and vowel recognition both in clean and noisy speech. This motivated a combination of feature streams, which improved scores compared to a recognizer using denoised MFCCs as feature input.

Acknowledgements

Supported by the DFG (SFB/TRR 31 'The active auditory system'; URL: http://www.uni-oldenburg.de/sfbr31). The OLLO speech database OLLO has been developed as part of the EU DIVINES Project IST-2002-002034.

We thank Thomas Brand, Michael Kleinschmidt, David Gelbart, Alexander Fischer, and Jörg-Hendrik Bach for their contributions to this work. We also thank Frank Ohl, Max Happel, and Arne Meyer for providing the physiological data shown in Fig. 1.

References


