Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition

Marc René Schädler[a] and Birger Kollmeier
Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, D-26111 Oldenburg, Germany

(Received 26 September 2014; revised 2 March 2015; accepted 5 March 2015)

To test if simultaneous spectral and temporal processing is required to extract robust features for automatic speech recognition (ASR), the robust spectro-temporal two-dimensional-Gabor filter bank (GBFB) front-end from Schädler, Meyer, and Kollmeier [J. Acoust. Soc. Am. **131**, 4134–4151 (2012)] was de-composed into a spectral one-dimensional-Gabor filter bank and a temporal one-dimensional-Gabor filter bank. A feature set that is extracted with these separate spectral and temporal modulation filter banks was introduced, the separate Gabor filter bank (SGBFB) features, and evaluated on the CHiME (Computational Hearing in Multisource Environments) keywords-in-noise recognition task. From the perspective of robust ASR, the results showed that spectral and temporal processing can be performed independently and are not required to interact with each other. Using SGBFB features permitted the signal-to-noise ratio (SNR) to be lowered by 1.2 dB while still performing as well as the GBFB-based reference system, which corresponds to a relative improvement of the word error rate by 12.8%. Additionally, the real time factor of the spectro-temporal processing could be reduced by more than an order of magnitude. Compared to human listeners, the SNR needed to be 13 dB higher when using Mel-frequency cepstral coefficient features, 11 dB higher when using GBFB features, and 9 dB higher when using SGBFB features to achieve the same recognition performance.

© *2015 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4916618]

[JFC] Pages: 2047–2059

I. INTRODUCTION

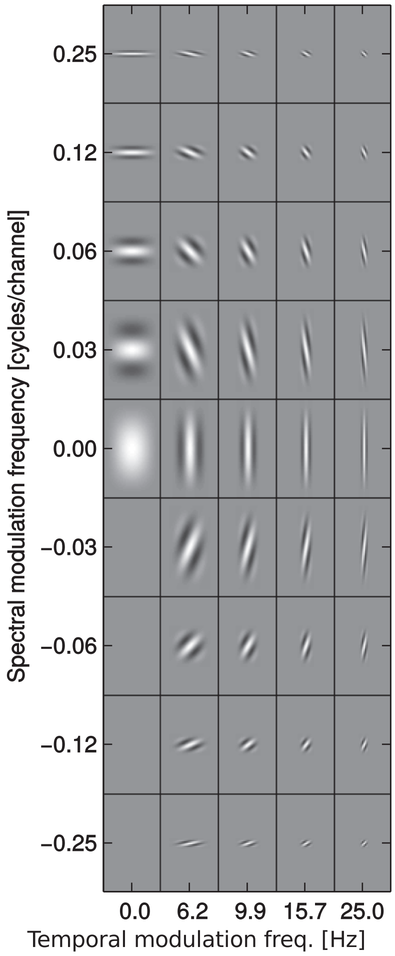
After years of investigation on robust automatic speech recognition (ASR), human listeners still outperform ASR systems in realistic acoustic environments (Lippmann, 1997; Meyer *et al.*, 2011; Barker *et al.*, 2013). Inspired by the ability of the human auditory system to decode speech signals in the most difficult acoustic conditions, many principles of auditory signal processing were integrated into ASR systems in attempts to improve their recognition performance. These approaches usually targeted the feature extraction stage (front-end), where the more tangible peripheral auditory processes can be mapped to signal processing algorithms and which is more specific to auditory processes than the recognition stage (back-end). The current study aimed to improve the front-end by extracting spectro-temporal modulation features with independent spectral and temporal processing instead of joint spectro-temporal processing.

Many of the speech representations (or features) used in ASR systems stem from spectro-temporal representations of sound that already incorporate basic auditory principles, such as the log Mel-spectrogram (LMSpec). The LMSpec is a spectrogram with a logarithmic amplitude and a Mel frequency scaling. It considers very basic auditory principles of the human auditory system, such as the resolution across frequencies and logarithmic perception of intensity. However, these static spectro-temporal representations themselves are not well suited as robust speech features because environmental changes, such as additive noise and reverberation, strongly affect them. The characteristics of the inherently dynamic speech signals are better represented in changes that occur in the spectro-temporal representations across frequencies and over time; this is why many robust features are extracted by encoding spectral or temporal *changes*. An example for spectral processing is the still widely used Mel-frequency cepstral coefficients (MFCCs), which perform a discrete cosine transform in the spectral dimension of a LMSpec (Davis and Mermelstein, 1980). An example for temporal processing is the calculation of discrete temporal first and second order derivatives, called deltas and double deltas, which are usually used to encode the dynamics of MFCC and other features. Many other, differently motivated spectral and temporal processing schemes were combined with the goal of improving the robustness of ASR systems (e.g., Hermansky, 1990; Hermansky *et al.*, 1992; Hermansky and Sharma, 1999; Nadeu *et al.*, 2001; Hermansky and Fousek, 2005; Moritz *et al.*, 2011) but without relating the spectral to the temporal processing nor vice versa.

In approaches to join spectral and temporal modulation processing, and thus allowing for higher order dependencies between both, Kleinschmidt (2002) and Kleinschmidt *et al.* (2002) found that the physiologically motivated (Qiu *et al.*, 2003) two-dimensional (2D) spectro-temporal Gabor filters were good candidates. Aside from their use in ASR systems, a number of studies suggested the use of 2D Gabor filters to extract spectro-temporal features for acoustic signal and speech analysis (e.g., Chi *et al.*, 2005; Mesgarani *et al.*, 2006;

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

Ezzat *et al.*, 2007). Because in early approaches to extract features with 2D Gabor filters the filter parameters were determined in a data driven way, and as a consequence some feature dimensions were highly correlated, Meyer and Kollmeier (2011) mapped these Gabor features to an intermediate phoneme probability layer by means of a tandem setup to use them with standard Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) based recognition back-ends. Recently, in an approach to structure the 2D Gabor filter parameter space and gain a set of universal 2D Gabor filters for robust speech recognition, the 2D Gabor filter bank (GBFB) features were introduced and shown to improve the robustness of ASR systems when they are used directly with standard GMM/HMM back-ends by Schädler *et al.* (2012) and Moritz *et al.* (2013). The 2D spectro-temporal filters of the GBFB, which were used to extract robust speech features by 2D-convolving each of them with a LMSpec, are depicted in Fig. 1 and cover a range of spectral and temporal modulation frequencies that were found to be beneficial for robust ASR. The extraction of GBFB features is explained in detail in Sec. II B. Meyer and Kollmeier (2011) attributed the improvements in robustness to a locally increased SNR due to the higher sensitivity to speech patterns of the more complex spectro-temporal patterns, most notably to the ability of discriminating upward and downward spectro-temporal patterns (cf. off-axis filters in Fig. 1). Schröder *et al.* (2013) found that using GBFB features can improve the recognition performance in a speech-unrelated acoustic event detection task; this confirms the universality of the GBFB filter set for acoustic recognition tasks. However, a model of joint spectro-temporal processing does not allow changes to the spectral processing without having an effect on the temporal processing and vice versa; this would imply that all models of separate spectral and temporal processing are insufficient. It is unknown to what extent spectral and temporal processing in the auditory system of mammals interact with each other (Depireux *et al.*, 2001; Qiu *et al.*, 2003). Further, the more complex 2D filtering process results in considerably higher computational costs for the feature extraction. If spectral and temporal processing were independent processes, the mentioned limitations would not apply.



Spectral modulation frequency [cycles/channel]
0.25
0.12
0.06
0.03
0.00
−0.03
−0.06
−0.12
−0.25
0.0
6.2
9.9
15.7
25.0
Temporal modulation freq. [Hz]

FIG. 1. Taken from Schädler *et al.* (2012). Filter shapes of the 2D Gabor filter bank (GBFB) filters. Each tile represents the filter function of a spectro-temporal 2D Gabor filter, where the horizontal axis within each tile is the temporal one and the vertical axis is the spectral one. The 2D filter functions are sorted by their spectral and temporal center modulation frequencies. To extract GBFB features, a LMSpec of speech is filtered by means of a 2D convolution with these filters. While the filters on the axis (0 Hz or 0 cycles/channel) are purely spectral or purely temporal filters and can be separated into a real-valued spectral 1D filter and a real-valued temporal 1D filter, the off-axis filters are inseparable spectro-temporal filters.

In this study, it was investigated whether the improvements in robustness gained with the structured, spectro-temporal GBFB approach require the complex joint 2D spectro-temporal processing or if a separate spectral and temporal processing with two 1D GBFB can be used to extract features that perform similarly or better. The basic idea was to replace the inseparable up- and downward 2D patterns of the GBFB with separable patterns and then perform the spectral and the temporal filtering separately with 1D Gabor filters. A 1D Gabor filter is depicted in Fig. 2 and the relation of 1D-spectral and 1D-temporal Gabor filters to the inseparable up- and downward 2D-spectro-temporal Gabor filters is illustrated in Fig. 3. In Fig. 3, it can be observed that the addition (A)/subtraction (S) of an inseparable 2D spectro-temporal downward (D) filter to/from its corresponding upward (U) filter is identical to the separable filter RR/II, which in turn can be described by a separate spectral and temporal filtering process with the real (R) or imaginary (I) part of 1D Gabor functions. The relation between a pair of a spectral and a temporal 1D filter, and the corresponding 2D filter is the outer product and is explained later in more detail. The combination of spectral and temporal filters with different phases, which were determined by the use of the real (R) or imaginary (I) part, but identical center modulation frequencies resulted in

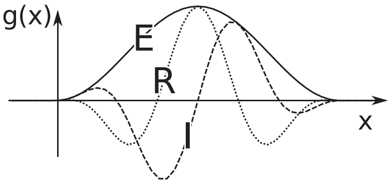


FIG. 2. Absolute (E), real (R), and imaginary (I) part of a complex-valued filter function of a 1D Gabor filter with 3.5 half-waves under the envelope. Each part is a real-valued function and can be used to filter a signal where R and I are band-pass filters with the same transfer function and only differ in the phase, while E describes a low-pass filter.

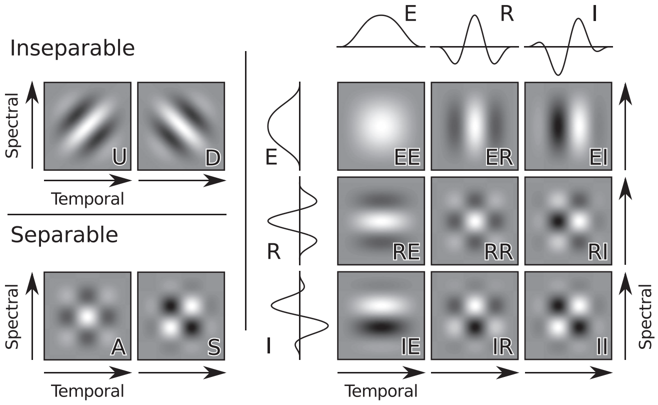


FIG. 3. Inseparable and separable 2D spectro-temporal Gabor filters and their relation to separate 1D spectral and 1D temporal Gabor filters. Each tile represents the filter function of a 2D spectro-temporal filter with the horizontal axis within each tile being the temporal and the vertical axis being the spectral one. Left panel: The 2D upward (U) and downward (D) filters are not separable, while their sum A (U + D) and difference S (U – D) are; right panel: Effective 2D filter shapes when applying subsequent spectral and temporal 1D filters using different parts of 1D Gabor filters (E, envelope; R, real part; I, imaginary part). The amplitude of the 2D filters is encoded in gray scale, where white means high amplitude and black low amplitude.

different effective spectro-temporal filter patterns (cf. RR, RI, IR, and II in Fig. 3). Hence each inseparable 2D filter in Fig. 1 could have been replaced with different separable 2D filters that have the same absolute spectral and temporal center modulation frequencies as the inseparable 2D filter.

Instead of only replacing the inseparable 2D filters, the whole 2D GBFB was replaced by two separate 1D GBFB: A spectral one and a temporal one. For these, the positive spectral and temporal center modulation frequencies were taken from the 2D GBFB. The phase of the employed filters was determined by taking the real (R) or the imaginary (I) part of the 1D Gabor filters. All spectral filters were assumed to have the same phase, and also all temporal filters were assumed to have the same phase, while spectral and temporal filters were allowed to have different phases. This structure allowed four SGBFB feature vectors with different combinations of spectral and temporal phases: Real-real (RR), real-imaginary (RI), imaginary-real (IR), and imaginary-imaginary (II) (cf. RR, RI, IR, and II in Fig. 3). To evaluate which of the phase combinations performs best in a robust ASR task, the four different SGBFB feature vectors were compared to GBFB and MFCC features on the CHiME (Computational Hearing in Multisource Environments) keyword recognition task. Barker *et al.* (2013) created the CHiME keyword recognition task to compare the robustness of ASR systems under controlled, realistic low-SNR conditions and to be able to compare the ASR performance to performance data from human listeners. Further, the role of the spectral and temporal modulation phase was assessed in recognition experiments combining several SGBFB feature vectors with different phase combinations.

II. METHODS

A. Spectro-temporal representation

The calculation of the LMSpec was based on an amplitude spectrogram with frames of 25 ms length and a temporal resolution of 100 frames/s. The linear frequency axis of the spectrogram was transformed to a Mel-scale using 31 equally spaced triangular filters with center frequencies in the range from 124 to 7284 Hz. The values of the amplitude Mel-spectrogram were subsequently converted to a decibel scale. All feature extraction schemes that are presented in the following extracted features from a LMSpec. An example of a LMSpec of a speech signal is depicted in the upper panel of Fig. 4.

B. Gabor filter bank features

2D GBFB features were extracted from a LMSpec using auditory-motivated spectro-temporal 2D Gabor filters, as described by Schädler *et al.* (2012). There a LMSpec was 2D convolved (filtered) with a set of 2D Gabor filters to model the response of a range of neurons in the auditory cortex to the presented spectro-temporal patterns. The 2D filter shapes that were used to extract GBFB features are depicted in Fig. 1. These filters were tuned to specific spectro-temporal modulation patterns that occur in speech signals and motivated by the fact that some neurons in the primary auditory cortex of mammals were found to be tuned to very similar spectro-temporal modulation patterns (Qiu *et al.*, 2003). A 2D Gabor filter represents an idealized spectro-temporal receptive field and requires a pairing of spectral and temporal modulation frequencies. The pair of modulation frequencies determines a filter's shape and, hence, which spectro-temporal pattern would yield the strongest response in this particular filter. The main parameters of the employed 2D Gabor filters were the spectro-temporal center modulation frequencies and the spectral and temporal modulation bandwidths. Schädler *et al.* (2012) structured the parameters of the 2D Gabor filters in a filter bank, which limited the number of free parameters and the correlation between the resulting feature dimensions. In this study, the

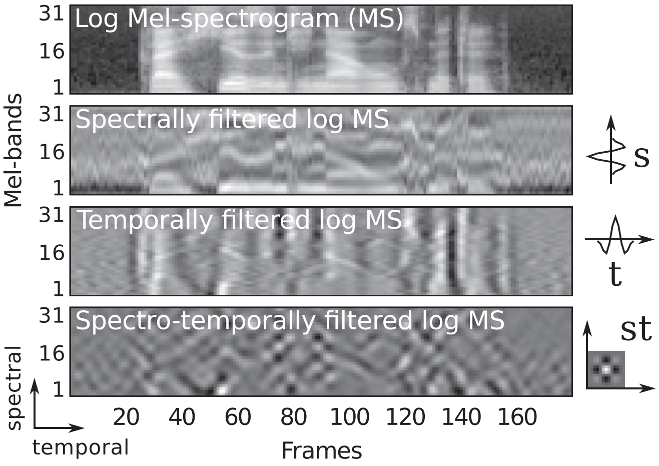


FIG. 4. Filtering a log Mel-spectrogram (LMSpec) by means of a 2D convolution: The LMSpec in the upper panel is 2D-convolved with a spectral 1D filter *s*, a temporal 1D filter *t* and the corresponding spectro-temporal 2D filter *st*. The result of the filtering process is depicted to the left of the corresponding filter. The amplitude of the 2D filters and (filtered) spectrograms is encoded in gray scale where white encodes high amplitude and black encodes low amplitude.

same set of GBFB parameters was used, which was optimized for ASR and confirmed to extract robust ASR features (Moritz *et al.*, 2013): The considered spectral modulation frequencies were $\omega_{\mathrm{s}} = 0.000$, 0.029, 0.060, 0.122, 0.250 cycles/channel. The considered temporal modulation frequencies were $\omega_{\mathrm{t}} = 0.0$, 6.2, 9.9, 15.7, 25.0 Hz. The number of half-waves under the envelope, which determines the bandwidth, in the spectral dimension was $\nu_{\mathrm{s}} = 3.5$. The number of half-waves under the envelope in the temporal dimension was $\nu_{\mathrm{t}} = 3.5$. The maximum extension of the filters in the spectral dimension was $b_{\mathrm{s}}^{\max} = 3 \cdot 31$, which is three times the number of Mel-bands. And the maximum extension of the filters in the temporal dimension was $b_{\mathrm{t}}^{\max} = 40$ frames (400 ms). The considered spectro-temporal center modulation frequencies were combinations of the spectral and temporal modulation frequencies and hence arranged on a grid (cf. Fig. 1). Spectral and temporal cross-sections through the maximum of the 2D frequency response of the GBFB filters with these parameters are shown in Fig. 5. To extract GBFB features from a LMSpec, it was convolved with each of the 41 2D Gabor filters, which resulted in 41 filtered LMSpecs. Subsequently, the filtered LMSpecs were spectrally sub-sampled at a rate of a quarter of the extent of the spectral width of the corresponding filter. This reduced redundancy from the filtered LMSpec, and was shown to be superior to using a Principle Component Analysis (Schädler *et al.*, 2012). The filtered and sub-sampled LMSpecs were concatenated and formed a 455-dimensional feature vector, which is referred to as GBFB features. The difference in dimensionality to the original GBFB features, which are 311-dimensional, was due to the larger bandwidth (8 vs 4 kHz) that was considered in this study.

C. Separate Gabor filter bank features

Separate Gabor filter bank features (SGBFB) were extracted with two 1D Gabor filter banks, one spectral and one temporal, instead of with a filter bank of 2D Gabor filters.

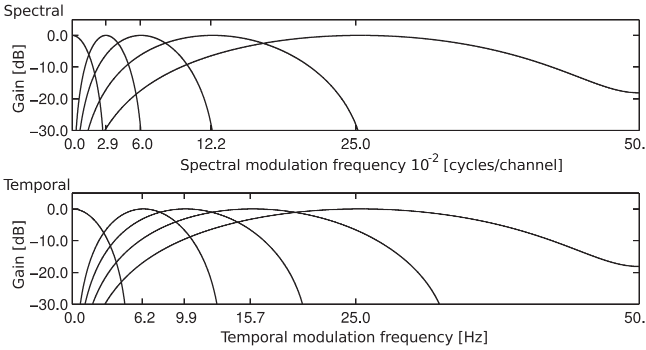


FIG. 5. Modified from Schädler *et al.* (2012). Upper panel: Spectral cross-sections through the maximum of the 2D frequency response of GBFB filters; Lower panel: Temporal cross-sections through the maximum of the 2D frequency response of GBFB filters. The overlap of adjacent band-pass modulation filters is constant and governed by the distance between them and by their bandwidth.

1. 1D Gabor filters

Equation (1) describes a 1D Gabor filter, where h_b is a Hann envelope function of width b, s_ω a sinusoid function with radian frequency ω, and g the product of both

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

$$s_\omega(x) = \exp(i\omega x), \qquad (1b)$$

$$g_{\omega,\nu}(x) = \underbrace{s_\omega(x)}_{\text{carrier}} \cdot \underbrace{h_{\nu/2\omega}(x)}_{\text{envelope}}. \qquad (1c)$$

The width b is inversely proportional to the radian frequency ω and proportional to the number of half-waves under the envelope ν. Consequently, all 1D Gabor filters $g_{\omega,\nu}$ with the same value for ν are constant-Q complex-valued band-pass filters, where ω is the (radian) center frequency and determines the scale of the filter. The complex-valued filter function of a 1D Gabor filter with $\nu = 3.5$ half-waves under the envelope is depicted in Fig. 2, where E marks the absolute values (or envelope), R the real part, and I the imaginary part of the filter. Each of the different parts (E, R, and I) can be used to filter a signal. While E describes a low-pass filter, R and I are band-pass filters that only differ in phase and share the same frequency response. The width b of the Gabor filters is limited by $b^{\max}$. Filters with $\omega = 0$ would have an infinitely large support, which is why in this case the width of the envelope is set to $b^{\max}$, effectively resulting in a low pass filter (E). These filters (E, R, and I) can be applied in the spectral or in the temporal dimension to a LMSpec, resulting in a spectral or temporal modulation filtering, respectively. In the following, a spectral filter bank and a temporal filter bank of 1D Gabor filters are presented.

2. 1D Gabor filter banks

The center modulation frequencies (ω), the maximum filter width $b^{\max}$, and the number of half-waves under the envelope ν, which determines the filters' Q-factor were taken from the GBFB [cf. parameters from Schädler *et al.* (2012)]. Hence the spectral modulation filter bank consisted of five 1D Gabor filters with $\nu = 3.5$, $b^{\max} = 93$ bands (three times the number of Mel-bands), and the following spectral modulation frequencies: $\omega = 0.000$, 0.029, 0.060, 0.122, and 0.250 cycles/band. The temporal modulation filter bank consisted of five 1D Gabor filters with $\nu = 3.5$, $b^{\max} = 40$ frames, and the following spectral modulation frequencies: $\omega = 0.0$, 6.2, 9.9, 15.7, and 25.0 Hz. As with GBFB filters, the envelope (E) function of width $b^{\max}$ was used as the filter function if the width of a filter function would exceed the maximum width $b^{\max}$, which here was the case for filters with $\omega = 0$. For all other filters ($\omega > 0$), only the real (R) or the imaginary (I) part of the filter was used as the filter function. As a result, in total, nine different spectral filters: 0.000 (E), 0.029 (R and I), 0.060 (R and I), 0.122 (R and I), and 0.250 (R and I) cycles/band, and nine different temporal filters: 0.0 (E), 6.2 (R and I), 9.9 (R and I), 15.7 (R and I), and 25.0 (R and I)

Hz were considered. The real (R) part and the corresponding imaginary (I) part only differed in phase and hence shared the same frequency response. As a consequence, the frequency responses of the 1D spectral and 1D temporal Gabor filters were exactly the same as the cross-sections through the maximum of the 2D frequency responses of the 2D GBFB filters depicted in Fig. 5. Hence the two 1D Gabor filter banks covered the same range of spectral and temporal modulation frequencies as the 2D Gabor filters of the GBFB.

3. 1D and 2D filtering of LMSpecs

The 1D filtering was performed by convolution with the corresponding filter functions. Temporal modulation filters were represented as row vectors and were convolved with each channel of the LMSpec independently. Likewise, spectral modulation filters were represented as column vectors and were convolved with each frame of the LMSpec independently. The temporal and spectral 1D filtering was performed by means of a 2D convolution with row and column vectors, respectively. Therefore the LMSpec was convolved with a 1D row or column vector, as defined in Eq. (2), where k and n are the spectral and temporal indices of the LMSpec, respectively, and i and j the spectral and temporal offset of the filter from its center, respectively,

$$\begin{aligned}&\text{filtered-LMSpec}(k,n)\\&\quad:=\sum_{i,j}\text{LMSpec}(k-i,n-j)\cdot\text{filterfunction}(i,j),\end{aligned} \tag{2}$$

filtered-LMSpec(k,n) was only calculated if LMSpec(k,n) existed, so that both the LMSpec and the filtered LMSpec, had the same size. In the following, a 2D convolution with a 1D filter, i.e., a filter the extent of which in the spectral dimension is one Mel-band or in the temporal dimension is one frame, is referred to as a 1D convolution or 1D filtering. Of course, a LMSpec can first be filtered spectrally, and the output can than be filtered temporally or vice versa. The order, i.e., if the spectral or temporal filtering is performed first, of this special form of spectro-temporal filtering does not affect the outcome. The outcome of a spectrally *and* temporally filtered LMSpec, is a spectro-temporally filtered LMSpec, and the corresponding spectro-temporal filter can be identified. In Eq. (3), a spectral filter s (column vector) and a temporal filter t (row vector) were applied in arbitrary order to a LMSpec

$$\begin{aligned}\text{filtered-LMSpec} &= [\text{LMSpec} * \text{s}] * \text{t} && \text{(3a)}\\ &= [\text{LMSpec} * \text{t}] * \text{s} && \text{(3b)}\\ &= \text{LMSpec} * \underbrace{[\text{s} * \text{t}]}_{\text{outer product: st}} && \text{(3c)}\\ &= \text{LMSpec} * \text{st}. && \text{(3d)}\end{aligned}$$

In Eq. (3c), the 1D convolution with s and t was identified as the 2D convolution with the outer product of s and t. Hence the outer product of a spectral 1D and a temporal 1D filter is a *separable* filter because it can be described by independent spectral and temporal filter operations. The same is true for any 2D filter that can be described by a separate spectral and temporal 1D filter. Figure 4 shows an example of a LMSpec of clean speech after filtering using temporal, spectral, and spectro-temporal filters. The corresponding filter functions are depicted to the right of the filtered LMSpecs.

4. Feature extraction

SGBFB features were extracted by first filtering the LMSpec spectrally, where either the R or the I phased filters were used, except for the DC filter ($\omega = 0$) for which always the E type was used. Due to the limited bandwidth in the output of spectral filtering processes with low center modulation frequencies, high correlations could be observed between some adjacent channels of the output. To the reduce these correlations, each spectrally filtered LMSpec was reduced in dimensionality by keeping only representative Mel-bands. This was achieved by critically sub-sampling the filtered LMSpec in spectral dimension at a rate of a quarter of the corresponding filters width b, where at least the center channel (Mel-band number 16), and at most all channels were kept. The same procedure for dimensionality reduction was used to extract GBFB features. The spectrally filtered and spectrally down-sampled LMSpecs were then filtered temporally, where either the R or the I phased filters were used, except for the DC filter ($\omega = 0$) for which always the E type was used. By the subsequent spectral and temporal filtering of the LMSpec, all considered spectral modulation frequencies were combined with all considered temporal modulation frequencies. The spectro-temporally filtered LMSpecs were concatenated and formed a 255-dimensional feature vector. These features are referred to as separate Gabor filter bank features or just SGBFB features.

With both the spectral and the temporal filter bank, the real (R) or the imaginary (I) part of the filters can be used. The filters that were actually employed are indicated by a suffix, where the first letter indicates the spectral and the second letter the temporal filter phase, e.g., SGBFB-RI. The effective spectro-temporal filter shapes for all possible combinations of all considered spectral and temporal E, R, and I filters are depicted in Fig. 6.

5. Spectro-temporal modulation phase

Because all four possible SGBFB feature vectors (SGBFB-RR, SGBFB-RI, SGBFB-IR, and SGBFB-II) covered the same range and combinations of spectral and temporal modulation frequencies and only differed in the phase of the modulation filters, it was investigated which phase combination offered the most robust representation in a speech-in-noise recognition experiment. Only two phase values were considered: The first one corresponded to the real (R) part of a Gabor filter (no phase shift), and the second one corresponded to the imaginary (I) part, where the carrier phase was shifted by $\pi/2$ rad relative to the real part. The real-real (RR) and imaginary-imaginary (II) spectro-temporal filters can be derived from the corresponding upward (U) and downward (D) filters by addition (A) and

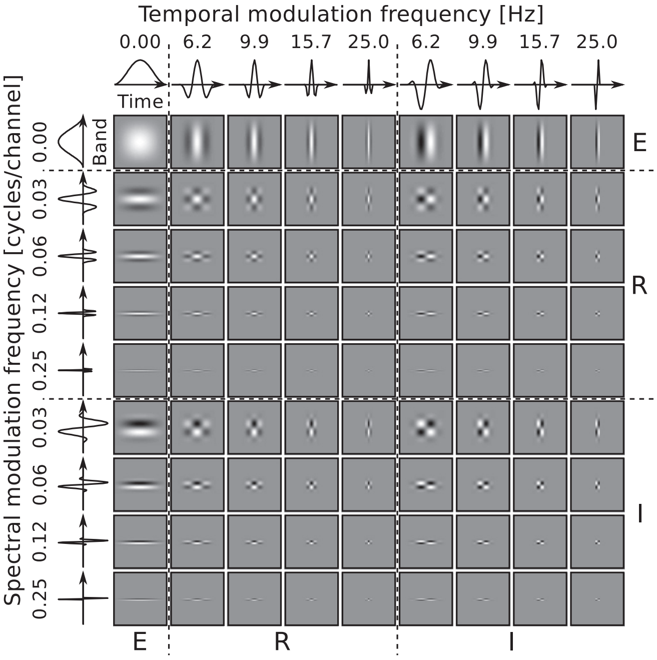


FIG. 6. All possible combinations of spectral and temporal 1D GBFB filters and their equivalent, separable spectro-temporal 2D filter functions. Each tile represents the outer product of the corresponding spectral and temporal filter functions with the horizontal axis within each tile being the temporal and the vertical axis being the spectral one. The 1D filters, depicted above and to the left of the 2D filters, are sorted by spectral and temporal center modulation frequencies, and are grouped according to the part of the complex 1D Gabor filter that is used: Envelope (E), real (R), imaginary (I). For a specific separate Gabor filter bank (SGBFB) feature vector, only a subset of these filters is used, which is indicated by a two-letter suffix. For example, for the SGBFB-RI feature set, the spectral E and R filters are combined with the temporal E and I filters. Note that each SGBFB feature vector covers the whole range of considered modulation frequencies, and that none of the 81 2D filter shapes is repeated.

subtraction (S) as depicted in Fig. 3, while the real-imaginary (RI) and imaginary-real (IR) phase combinations cannot be represented by linear combination of any two filters of the 2D GBFB. To take multiple phase combinations into account, different *single* SGBFB feature vectors were concatenated and the robustness of the combined—or *dual*—SGBFB feature vectors was determined in a speech-in-noise recognition experiment. The concatenation of two 255-dimensional, single feature vectors resulted in a 510-dimensional dual feature vector and is referred to as SGBFB-X-Y, where X determined the phases of the first and Y the phases of the second vector, e.g., SGBFB-RR-II. A dual SGBFB feature vector represented all spectro-temporal modulation frequencies twice, in contrast to the 455-dimensional GBFB feature vector, where only the modulation frequencies of the truly spectro-temporal filters were represented twice [cf. upward (U) and downward (D) filters in Fig. 1]. This explains the difference in dimensionality between dual SGBFB feature vectors and the GBFB feature vector. The concatenation of feature vectors with all possible phase combinations combined all considered spectral and temporal 1D filters and hence extracted 1020-dimensional feature vectors effectively using all 81 2D patterns depicted in Fig. 1. These feature vectors are referred to as *complete* SGBFB features or SGBFB-RR-RI-IR-II.

D. Feature normalization

Blind feature statistics adaptation, such as mean and variance normalization (MVN) (Viikki and Laurila, 1998) or histogram equalization (HEQ) (De la Torre *et al.*, 2005) can improve the robustness of an ASR system. All features were normalized using histogram equalization (HEQ). As each feature dimension was processed independently, the process is only described for one feature dimension, which is considered to be a time series. While mean and variance normalization normalizes the first two statistical moments of the distribution of the values of the time series, HEQ can normalize even higher statistical moments, such as skewness and kurtosis. For this, the values of the time series were projected by a function that mapped the source distribution to a desired target distribution. The mapping function was estimated by calculating 100 percentiles (e.g., 0.5%, 1.5%, …, 99.5%) of the source distribution and mapping these to the same percentiles of the desired target distribution, where values between the percentiles were interpolated linearly. Care needed to be taken when estimating the percentiles of the source distribution, as the 0% and 100% percentiles could not be reached with finite time series. The maximum expected percentile $\mathrm{p}_N^{\max}$ and minimum expected percentile $\mathrm{p}_N^{\min}$ when drawing N samples from a distribution were estimated using Eq. (4),

$$\mathrm{p}_N^{\max} = 100 * \frac{N}{N+1}, \tag{4a}$$

$$\mathrm{p}_N^{\min} = 100 * \frac{1}{N+1}. \tag{4b}$$

Therefore, 100 equally spaced percentiles between $\mathrm{p}_N^{\min}$ and $\mathrm{p}_N^{\max}$ were mapped to the corresponding percentiles of the standard normal distribution, where N was the number of feature vectors. The resulting time series had—within the limits due to mapping only 100 percentiles—the same moments as the standard normal distribution. All features were processed with HEQ on a per-utterance basis, where the average utterances length of the employed corpus was 1.8 ± 0.25 s.

E. Recognition experiment

The task that was employed to evaluate the robustness of ASR systems is the recognition of English commands being spoken in noisy living room environments that were recorded using an binaural manikin. Therefore the training, development, and test data sets from the first track of the second CHiME challenge (Vincent *et al.*, 2013) were used. The sentences of this corpus were recorded from 34 different (male and female) speakers. They have a fixed syntax of the form “command color preposition letter number adverb” (e.g., “put red at G9 now”), where the words were drawn from a closed vocabulary. The utterances of the development and test data set were filtered with the binaural combined head and room impulse responses of two rooms (a lounge and a kitchen) corresponding to a frontal position at a distance of 2 m. Subsequently, they were mixed with noise samples recorded using the binaural manikin in the same environments at SNRs from −6 to 9 dB. In this study, the binaural signals were mixed down to one channel prior to the feature extraction by adding the left and the right channel. The whole sentences had to be recognized but only the percent correct value of the letter (in the example: G) and the digit (in the example: 9) was evaluated as in the first track of the second CHiME challenge.

Three different training data sets were available and used to evaluate the performance of ASR systems depending on the training condition: Clean, reverberated, and isolated (which is noisy and reverberated). While the clean data set contained unprocessed speech samples, the utterances of the reverberated and isolated data sets were filtered with the binaural impulse responses. The utterances of the isolated (or noisy) data set were additionally mixed with noise samples that were recorded with the binaural manikin in the corresponding room at SNRs from −6 to 9 dB. Even though some of the considered front-ends might have performed better with additional training data, the unmodified training data sets from the CHiME challenge were used for the sake of comparability. For evaluation, each ASR system was trained with the three different training data sets. While all pilot experiments had been conducted with the development data set, the results were obtained on the test data set.

The training and testing scripts provided in the CHiME challenge are based on HTK (Young *et al.*, 2009). The differences between the provided scripts and the scripts that were actually used for conducting the experiments are highlighted in Sec. II G. For each training data set, the recognition performance in percent-of-digits-and-letters correct was measured at SNRs from −6 to 9 dB in 3 dB steps. The uncertainty of the performance measure due to the limited amount of test sentences (600) was estimated in advance, because it consisted of 1200 independent binary decisions; 600 for digits and 600 for letters. At 50% correct it happened to be about 1.45 percentage points, at 70% correct about 1.32 percentage points, and at 90% correct it was estimated to be about 0.85 percentage points. The recognition results, which depend on the SNR, were compared between different systems by calculating the relative change in SNR that would be required to get the same performance with two different systems, as described in Sec. II F. Additionally, human recognition performance data from the first CHiME challenge was available and used to present selected results in terms of the remaining *man-machine gap,* as described in Sec. II H.

F. Robustness measure

To report the relative improvement of a system over a reference system in a single value with physical meaning, the equal-performance increase in dB SNR (EPSI) is reported. This type of reporting is related to the speech reception threshold, which is widely used to measure the performance of human listeners to recognize speech in noise. The speech reception threshold is the SNR that is required to understand a specific portion, e.g., 50%, of the presented speech material. To use all available data points, the comparison was carried out at different performance levels. Hence the difference in SNR between the performances of two recognition system was integrated over the performance range where two systems could be compared. Let $P(r)$ be the performance graph of an ASR system, with r being the SNR in dB and P being the recognition performance at that SNR. Applying Eq. (5) guarantees the monotonicity of the performance graphs $P^{mon}(r)$,

$$P^{mon}(\mathrm{SNR}) = \min_{r \geq \mathrm{SNR}} P(r). \qquad (5)$$

The performance levels at which the systems were compared were interpolated in 0.5 dB steps in the region that data for both systems was available, as illustrated in Fig. 7. The average over the differences in SNR is invariant under any monotonic transformation of the performance axis. It is intuitively interpreted as the increase (or decrease) in SNR that is needed to get the same performance with the compared system as with the reference system. When comparing two ASR systems A and B, a symmetric EPSI was achieved by averaging the differences with A as the reference for B and with B as the reference for A. Ideally, the recognition performance of human normal-hearing listeners would have been used as a reference for all experiments. Although human performance data existed for the employed task, the human speech recognition (HSR) performance at the lowest SNR (−6 dB) was about 90% word recognition rate; so good that only few ASR systems could have been compared to it. Hence a reference ASR system was used instead, and only the best performing systems were compared to HSR performance.

G. Reference systems

Standard MFCCs and GBFB features with HEQ served as standard reference features. MFCCs were extracted from a LMSpec by spectrally processing it with a discrete cosine transform, where only the first 18 coefficients, which account

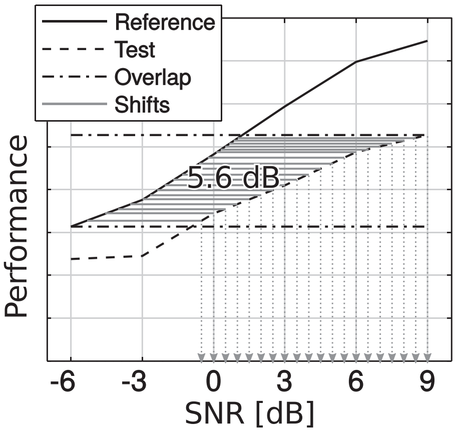


FIG. 7. Illustration of comparing the robustness of two ASR systems in terms of changes in signal-to-noise ratio (SNR). The average relative increase/shift of the SNR for a test ASR system that is required to achieve equal performance with a reference system can be calculated independently from the scaling of the performance axis. Therefore the integration points are selected on the SNR axis in 0.5 dB steps in the range where the performance graphs overlap on the performance axis.

for spectral modulation frequencies from 0 to 0.29 cycles/channel, were used. The 18 MFCCs were concatenated with their first and second discrete temporal derivatives, which were calculated by applying a temporal slope filter of five frames length once or twice, respectively. The resulting MFCC feature vector, which included both derivatives, was 54-dimensional. The extraction of the 455-dimensional GBFB feature vectors is described in detail in Sec. II B. All features that were evaluated in this study were normalized using HEQ as described in Sec. II D.

On the back-end side, GMM and HMM were used to model speech. The training and testing scripts provided in the first CHiME challenge (Barker *et al.*, 2013) are based on the Hidden Markov Toolkit (HTK) (Young *et al.*, 2009). Deviating from the default configuration, the reference system used tri-phone models instead of whole-word models. The required changes to the training procedure were based on the *HTK Wall Street Journal Training Recipe* from Vertanen (2006). Three-state left-to-right tri-phone speaker-depended acoustic models, a three-state background model with skip and back transitions, and a one-state short pause model tied to the center state of the background model were employed. The CMU Pronouncing Dictionary (Weide and Rudnicky, 2008), version 7a, was employed to generate initial monophone labels, where an optional short pause was allowed between two words. After the initial training of speaker-independent monophone models, tri-phone models of all possible monophone combinations were generated and initialized with the model of the center monophone. The parameters of the tri-phones were re-estimated in four iterations and subsequently tied with tri-phones that share the same center monophone using HTK's tree-based state tying method. The decision tree phonetic questions that are needed for the tree-based state tying were taken from Vertanen (2006). The threshold that governs the number of tied states was chosen so that the number of tied states was 700 ± 2. The number of Gaussian mixture components per state was increased stepwise to 2, 3, 5, and 7 in the course of the training procedure, with four iterations of parameter re-estimation in-between. The models were then adapted to the speaker using HTK's maximum *a posteriori* (MAP) method to update the mean values and the mixture weights, instead of using HTK's parameter re-estimation. The recognition of utterances was performed with the corresponding speaker-dependent model, where a language model enforced the syntax of recognized sentences (command color preposition letter number adverb).

H. Man-machine gap

To put the results of this study into the perspective of building an ASR system that is as robust as a normal-hearing human listener, selected results were compared to literature data of HSR performance, which is available from the first CHiME challenge (Barker *et al.*, 2013). The difference between the first CHiME challenge and the first track of the second CHiME challenge is that in the latter head movements of the speaker are simulated, which we consider to have a negligible effect on the HSR data for our purposes. The equal-performance increase in dB SNR (EPSI) of the ASR over the HSR results was used to quantify the remaining *man-machine gap*. In addition, the results for a GBFB-based system from the literature, which was presented by Moritz *et al.* (2013) during the second CHiME keyword recognition challenge and placed second, were also compared. This system, referred to as GBFB-CC, exploited binaural information using source separation based on non-negative matrix factorization, and featured a more sophisticated speaker adaptation, which includes in addition a maximum likelihood linear regression (MLLR) parameter adaptation step.

I. Reference implementations

MATLAB reference implementations of several methods, including the calculation of the LMSpec, MFCC features, GBFB features, SGBFB features, the HEQ, and the EPSI, are available online (Schädler, 2014).

III. RESULTS

All evaluated features sets were normalized using HEQ, as described in Sec. II D, and evaluated on the CHiME keyword-in-noise recognition task, as described in Sec. II E, using the ASR system described in Sec. II G, where the reference features were replaced with the features in question. The relative improvements are reported in EPSI, which is defined in Sec. II F. The uncertainty of all results was propagated from the estimated uncertainty due to the limited number of test sentences, as explained in Sec. II E.

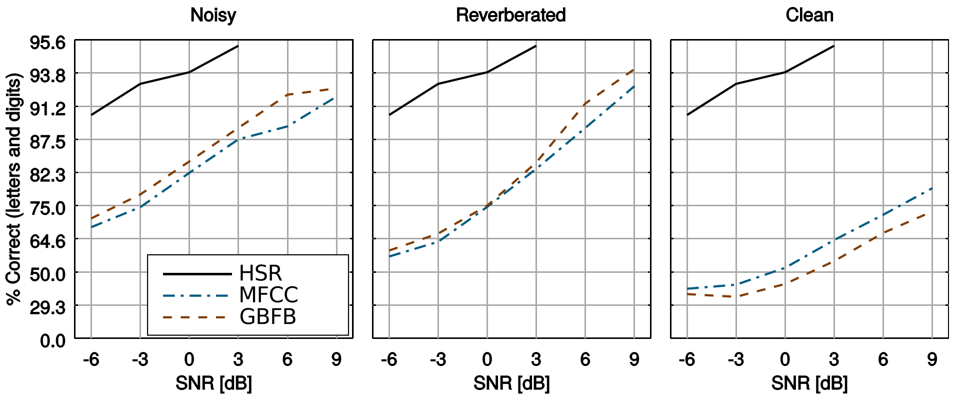


FIG. 8. (Color online) Recognition performance on the test data set of the MFCC and GBFB-based reference ASR systems depending on the SNR and the training data set, along with the approximate human speech recognition (HSR) performance. The word recognition rate in percent correct is plotted over the test SNR for systems trained with *clean, reverberated,* and *noisy* speech data. The *y* axis is a logarithmically scaled word error rate axis, which is labeled with the corresponding word correct rates in percent.

TABLE I. Recognition performance of the MFCC and GBFB-based reference ASR systems on the second CHiME keyword-in-noise recognition task in percent correct along with the human speech recognition (HSR) performance, which was measured during the first CHiME challenge. The systems were trained with clean, reverberated, or noisy data, and evaluated on the noisy test data set.

Features	Train condition	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB
HSR	—	90.3	93.0	93.8	95.3	96.8	98.8
MFCC	Clean	40.3	42.8	52.1	64.2	72.5	79.2
MFCC	Reverberated	57.4	63.5	74.7	83.0	88.9	92.8
MFCC	Noisy	68.7	74.6	82.2	87.5	89.1	92.0
GBFB	Clean	36.9	35.1	43.2	55.3	66.8	73.4
GBFB	Reverberated	60.0	66.5	75.0	84.1	91.4	94.0
GBFB	Noisy	71.4	77.8	84.2	88.9	92.2	92.7

A. Performance of reference system and data representation

The absolute recognition scores of the reference systems along with the approximate HSR performance depending on the SNR in decibels are depicted in Fig. 8, and reported in numerical form in Table I. As expected, the human performance was found to be superior to the performance of the ASR systems. Independent of the used features, the ASR systems that were trained on the noisy data set performed better at lower SNRs (less than 3 dB), while for high SNRs, particularly at 9 dB SNR, the systems trained on only reverberated data performed better. The ASR systems that were trained on the clean data set performed much worse, which is why these results were not considered to be a good indicator for robustness. Because we were interested in noise robustness, not the ability of generalizing from quiet to noisy conditions, the results for ASR systems trained with noisy data were taken as the indicator of robustness. To compare the ASR system with different features regarding their robustness on the CHiME task, the EPSI measure presented in Sec. II F was used to report the difference in performance in a single, physically interpretable value; the equal-performance increase of the SNR in decibels. The EPSIs of the reference ASR systems over the HSR performance in dB are reported in Table II. The MFCC-based reference system required the SNR to be +13.2 ± 0.95 dB higher to perform as well as an average native human listener, while the GBFB-based reference system required the SNR only to be +10.6 ± 1.12 dB higher. Hence the GBFB-based system was found to be more robust than the MFCC based system on this task. For ASR systems that do not reach HSR performance, such as the systems trained with clean data, the EPSI over HSR performance cannot be calculated. This is the reason why in the following the GBFB-based reference system was used as the baseline for the comparison.

TABLE II. Equal-performance increase in dB SNR (EPSI) of the reference ASR systems over HSR data for different training conditions, where a value of X means the SNR needs to be increased by X on average for the corresponding system to perform as well as a human listener. Using GBFB features reduced the distance to human performance compared to when using MFCCs from 13.2 dB to 10.6 dB SNR.

Features	Noisy	Reverberated	Clean
MFCC	+13.2 ± 0.95	+12.6 ± 1.00	—
GBFB	+10.6 ± 1.12	+10.3 ± 1.06	—

TABLE III. Average equal-performance increase in dB SNR over the GBFB reference system to achieve the same performance with single SGBFB features when training on clean, reverberated, or noisy data. A positive value indicates that the system under consideration performs worse than the GBFB reference system.

Features	Noisy	Reverberated	Clean
SGBFB-RR	+2.2 ± 0.45	+0.7 ± 0.30	−0.7 ± 0.36
SGBFB-RI	+2.7 ± 0.45	+0.8 ± 0.27	−1.1 ± 0.31
SGBFB-IR	+1.1 ± 0.44	+1.5 ± 0.26	−0.0 ± 0.30
SGBFB-II	+2.5 ± 0.42	+1.7 ± 0.28	−0.9 ± 0.35

B. Single SGBFB features

Table III reports the EPSIs of the differently phased 255-dimensonal SGBFB features over the GBFB reference system. We considered the results for the noisy training condition to carry the most information about the features' ability to facilitate the back-end of the recognition of speech in noise. The relative increase in SNR to achieve equal performance for the clean and reverberated training condition are reported for completeness. The SGBFB-IR system, which uses the imaginary part of the 1D Gabor filter for spectral filtering and the real part for temporal filtering, is the one that came closest to the GBFB reference with a EPSI of +1.1 ± 0.44 dB. This means that the ASR system with SGBFB-IR features required the SNR to be +1.1 ± 0.44 dB higher than with GBFB features to get the same performance. With the other SGBFB features, the EPSI increased to more than 2 dB. The ASR system with GBFB features outperformed all ASR systems using only single SGBFB feature vectors or MFCCs.

C. Dual SGBFB features

The required increase in dB SNR for all ASR systems using dual SGBFB feature vectors, which are combinations of two differently phased single SGBFB feature vectors, to achieve equal performance with the GBFB reference system are reported Table IV. The best dual SGBFB feature set was the one that concatenates SGBFB-RI and SGBFB-IR feature vectors to 510-dimensional SGBFB-RI-IR feature vectors. It yielded an improvement over the GBFB reference of −0.9 ± 0.45 dB, i.e., a decrease in SNR to achieve the same

TABLE IV. Average equal-performance increase in dB SNR over the GBFB reference system for ASR systems with MFCC or dual SGBFB features when being trained on clean, reverberated or noisy data.

System	Noisy	Reverb	Clean
SGBFB-RR-RI	−0.3 ± 0.46	+0.4 ± 0.28	+0.7 ± 0.30
SGBFB-RR-IR	+1.2 ± 0.42	+0.7 ± 0.28	−0.1 ± 0.34
SGBFB-RR-II	−0.7 ± 0.47	−0.0 ± 0.29	−0.5 ± 0.31
SGBFB-RI-IR	−0.9 ± 0.45	+0.1 ± 0.29	−1.0 ± 0.35
SGBFB-RI-II	+1.8 ± 0.43	+0.7 ± 0.28	−1.7 ± 0.31
SGBFB-IR-II	−0.4 ± 0.43	+0.6 ± 0.28	−1.0 ± 0.36

performance. In terms of word error rates, this translates to an average relative improvement of 8.3% over the GBFB reference system, and 20.6% over the MFCC reference system, where an improvement of 50% would correspond to halving the word error rate. The dual SGBFB feature vectors with the *same temporal phase* and different spectral phases (R**R**-I**R**, R**I**-I**I**) performed worse than the GBFB reference. Those with the same spectral phase and *different temporal phases* (I**R**-I**I**, R**R**-R**I**) performed as well as GBFB features within the uncertainty imposed by the setup. Those with *different spectral and temporal phases* (**RI-IR**, **RR-II**) improved the robustness of the GBFB-based reference system.

Using the MATLAB reference implementation, the 2D GBFB spectro-temporal filtering achieved a real-time factor of 0.4887 (median of 100 runs), while the 1D SGBFB-RI-IR spectro-temporal filtering achieved a real time factor of 0.0078 (median of 100 runs) on the same PC system,[1] i.e., the separate processing was found to be about 60 times faster. Hence by using dual SGBFB features instead of GBFB features, the computational time required for the spectro-temporal filtering was reduced by more than an order of magnitude, while at the same time the robustness was increased.

D. Complete SGBFB features

When concatenating all differently phased SGBFB features to 1020-dimensional SGBFB-RR-RI-IR-II feature vectors, the EPSI over the GBFB reference was -1.2 ± 0.42 dB when training on noisy data. In terms of word error rates, this translates to an average relative improvement of 12.8% over the GBFB reference system, and 24.8% over the MFCC reference system, where 50% would mean halving the word error rate. The most robust front-end evaluated in this study was the complete SGBFB feature set.

E. Quantity of training data

A reasonable question when using ASR systems with high-dimensional features is whether sufficient training data are available because the number of GMM parameters increases proportionally with the number of feature dimensions. On the one hand, using scarce training data could favor systems that require less parameters to be determined during the training phase and prevent systems with more parameters from showing their full potential. On the other hand, using large amounts of training data could conceal the possibility that systems using high-dimensional features might *require* these amounts of data, while systems with low-dimensional features would not perform worse when using less training data.

To test if one or the other was the case, systems with the low-dimensional MFCC features and the high-dimensional SGBFB-RI-IR features were trained with a reduced training data set, which contained only half of the training sentences that were available per speaker, i.e., 250 instead of 500. With this reduced training data set, the system that uses the 54-dimensional MFCC features performed 2.2 ± 0.44 dB (EPSI) worse and the system that uses the 510-dimensional SGBFB-RI-IR features performed 2.0 ± 0.46 dB worse compared to when using the full training data set. This result shows that the systems with high- and low-dimensional features were equally affected when the amount of training data was halved, and hence that no system was favored due to the amount of training data that were used in the recognition experiments. Compared to the system with MFCC features that was trained on the full training data set, the system with SGBFB-RI-IR features that was trained with the reduced training data set performed about (± 0.5 dB) the same. Hence we are confident that the training data set from the CHiME challenge provided a fair comparison of the differently-dimensional feature sets.

F. Remaining man-machine gap

Figure 9 depicts the absolute word recognition rates of the reference systems, the best SGBFB system, the GBFB-CC system, and from HSR experiments. Table V reports the EPSIs over human speech recognition performance that quantify the remaining *man-machine gap*. While the MFCC-based reference ASR system required the SNR to be about 13 dB higher to perform as well as a human listener, the GBFB-based reference system still had an EPSI of about 11 dB, and the best SGBFB-based system one of about 9 dB. Hence the gap in speech recognition robustness between man and machine remains but was reduced by 2 dB by using SGBFB features instead of GBFB features.

IV. DISCUSSION

A. Modulation phases

The main results reported in Tables III and IV indicate that an ASR system with a combination of SGBFB features may exhibit a greater robustness than the GBFB reference system if the phase of the spectral and temporal modulation filters is chosen in an appropriate way. The ASR systems with single SGBFB features vectors (RR, RI, IR, and II), which consider only one spectral and one temporal phase constellation, were found to be less robust than the GBFB reference system, where the systems with real-phase temporal filters (I**R** and R**R**) performed better than those with imaginary-phase temporal filters (I**I** and R**I**). To build a system with SGBFB features that was at least as robust as the reference system with GBFB features, a dual SGBFB feature vector with both temporal phase constellations was required (R**R**-R**I**, R**R**-I**I**, R**I**-I**R**, and I**R**-I**I**). If the temporal phase was the same (R**R**-I**R** or R**I**-I**I**), the corresponding system performed worse than the GBFB reference system. To improve the robustness of the GBFB reference system, both temporal and both spectral phase constellations were required (**RR-II** and **RI-IR**). Finally, the ASR system using complete SGBFB features, which include all possible phase combinations (RR-RI-IR-II), was found to be the most robust one. These findings suggest that the temporal phase is more important than the spectral phase and that diverse phase information of modulation filters is beneficial to the robustness of ASR systems.

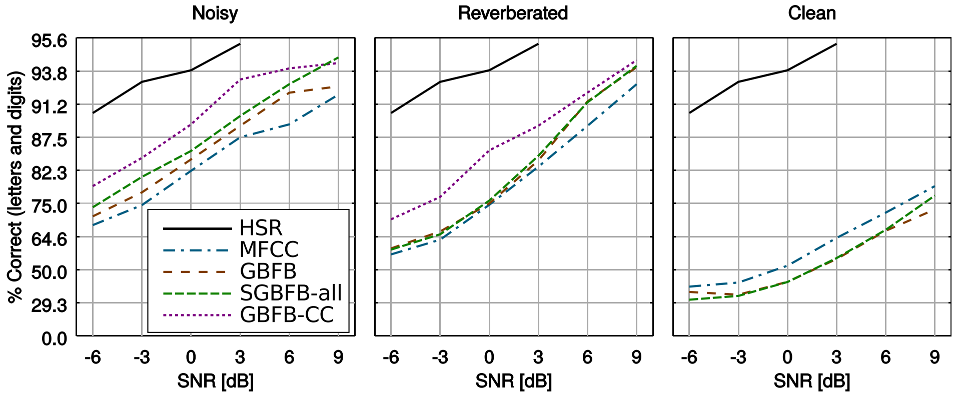


FIG. 9. (Color online) Recognition performance on the test data set of different ASR systems and human speech recognition (HSR) experiments depending on the SNR and training data set. Besides the performance of ASR systems using MFCC features, GBFB features, or the complete SGBFB feature set (SGBFB-all), the performance of a GBFB-based system with binaural processing (GBFB-CC) from Moritz *et al.* (2013) from the second CHiME challenge is depicted. The word recognition rate in percent is plotted over the test SNR. The *y* axis is a logarithmically scaled word error axis, which is labeled with the corresponding word correct rates in percent correct.

The phase of the modulation filters was found to be an important factor. However, it does not affect the frequency response of the filters, which indicates that modulation filters in the context of robust ASR are insufficiently described by only specifying their frequency response. A reason that considering temporal and spectral modulation filters with orthogonal, shifted carrier functions (i.e., the real and the imaginary part) benefits the robustness of ASR systems could be that their output is not systematically correlated, which is a property that GMMs with diagonal covariance matrices are well-disposed to. For example, for the temporal domain, the shape of the imaginary filter (I) is very similar to the shape of the slope (or delta) filter, which is traditionally used to calculate the first discrete temporal derivative with MFCCs, and the shape of the real filter (R) is very similar to the shape of the double-delta filter, which is traditionally used to calculate the second temporal derivative. Both describe different properties and seem to encode complementary information, which is why the combination of differently-phased feature vectors could improve the robustness. But while with the delta filters only one temporal center frequency was extracted, with the SGBFB filters considered here, five different modulation frequencies between 0 and 25 Hz were extracted.

TABLE V. Equal-performance increase in SNR over HSR performance in dB for different training conditions, where a value of X means the SNR needs to be increased by X on average for the corresponding system to perform as well as a human listener. Using GBFB features reduces the distance to human performance compared to when using MFCCs from 13.2 to 10.6 dB SNR. The use of dual SGBFB features can reduce the distance to 9.5 dB, and the use of all SGBFB feature vectors combined can reduce the distance further to 8.6 dB. The GBFB-based system from Moritz *et al.* (2013) which, like humans and opposed to the other systems, exploits binaural information (GBFB-CC), even gets as near as 6.2 dB to human performance.

System	Isolated	Reverb	Clean
MFCC	+13.2 ± 0.95	+12.6 ± 1.00	—
GBFB	+10.6 ± 1.12	+10.3 ± 1.06	—
GBFB-CC	+6.2 ± 1.19	+9.4 ± 1.04	—
SGBFB-RI-IR	+9.5 ± 1.19	+10.1 ± 1.04	—
SGBFB-RR-RI-IR-II	+8.6 ± 1.01	+10.2 ± 1.09	—

While the whole spectral context was always available to the back-end in the same feature vector, the temporal context was distributed over several feature vectors. A reason why the temporal phase was found to be more important than the spectral phase in this regard could be that the HMM back-end is inherently probabilistic about *timing* and could have benefited from the presence of additional hard-coded temporal information in the feature vectors. The effect of changing the temporal phase is that the carrier is shifted in time, while the window function (the envelope) remains invariant. The output of the temporal filters with shifted carriers could have conveyed information that otherwise was not accessible to the back-end.

B. 1D vs 2D Gabor filter complexity

Separating the spectro-temporal 2D GBFB into two SGBFB was not only found to improve the robustness of an ASR system in difficult acoustic conditions but also to achieve this with less complex filters. While with the 2D GBFB filters the spectral filtering and the temporal filtering are dependent and happen simultaneously, with the 1D SGBFB filters, the spectral and the temporal filtering are independent and can be carried out in arbitrary order. This reduces the complexity of the features and also of the feature calculation because no spectro-temporal interactions need to be considered. The corresponding reduction in computational time, that was required for the spectro-temporal processing, was found to be more than an order of magnitude. It is yet to be investigated if the 1D Gabor filters and the chosen parameter values for the filter width and center modulation frequencies are the optimal choice for robust ASR. But, at least in the studied context, it seems that truly spectro-temporal filters did not give an advantage over separate spectro-temporal filters. This suggests that future research on robust speech features might reasonably assume spectro-temporal interactions (such as, e.g., temporal changes of spectral information as in glides or formant transitions) to play a minor role in comparison to having both temporal and spectral information available simultaneously.

C. Remaining man-machine gap

A part of the remaining gap between the complete SGBFB feature based system (SGBFB-RR-RI-IR-II) and the HSR performance in Table V could be due to the very basic binaural processing (down-mixing) that was employed in this study, which did not exploit binaural cues for noise reduction as opposed to the human auditory system and the GBFB based system from the chime challenge (GBFB-CC). A SGBFB based system that exploits binaural information could provide further improvements in robustness. Another—maybe even related—reason could be the negligence of any phase—not modulation phase—information of the spectral channels. The temporal fine structure, which encodes binaural information as well as information about voicing or the harmonic structure of a signal, is not considered at all when using a LMSpec as a basis for feature extraction. This information could help to group signal parts and better separate them from the rest. The current research on this topic in the field of computational acoustic scene analysis (CASA) might some day converge with the investigation on robust speech recognition. For now, the SGBFB feature extraction algorithm permits the investigation of spectral and temporal modulation processing independently and to assess the interdependence of both types of processing in the context of speech recognition.

Even though the omission of certain modulation frequencies or spectro-temporal modulation pairs might be a good tool to systematically evaluate the relative importance of these features, this endeavor was beyond the scope of this paper and might be considered in future work.

V. CONCLUSIONS

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

(1) A combination of separate spectral and temporal 1D Gabor modulation filter banks (SGBFB) was successfully employed instead of the spectro-temporal 2D GBFB to extract robust ASR features. SGBFB features improved the robustness over GBFB features by up to 1.2 dB SNR, which corresponds to an average relative improvement of the word error rate of 12.8% over a GBFB based reference system, and 24.8% over a MFCC based reference system.

(2) While a close interaction between temporal and spectral processing was found to be comparatively irrelevant for robust ASR, the *phase* of the spectral and especially the temporal modulation filters was found to be an important factor, which can be used to provide complementary and additional temporal information to the back-end.

(3) Compared to human listeners, the SNR needed to be 13 dB higher for a MFCC-based system, 11 dB higher for a GBFB-based, and 9 dB higher for a SGBFB-based system, to achieve the same recognition performance.

ACKNOWLEDGMENTS

This work was supported by the DFG (SFB/TRR 31 "The active auditory system"). We would like to thank Bernd Meyer for his support and contribution to this work.

[1] CPU: AMD A10 PRO-7350B @ 2.1 GHz; RAM: DDR3L-1600 CL9; MATLAB R2010b (single thread) on LINUX 3.13.

Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P. (**2013**). "The PASCAL CHiME speech separation and recognition challenge," Comput. Speech Lang. **27**, 621–633.

Chi, T., Ru, P., and Shamma, S. A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. **118**, 887–906.

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

De la Torre, A., Peinado, A. M., Segura, J. C., Pérez-Córdoba, J. L., Benítez, M. C., and Rubio, A. J. (**2005**). "Histogram equalization of speech representation for robust speech recognition," IEEE Trans. Speech Audio Process. **13**, 355–366.

Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. (**2001**). "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," J. Neurophysiol. **85**, 1220–1234.

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

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

Hermansky, H., and Fousek, P. (**2005**). "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proceedings of Interspeech 2005*, pp. 361–364.

Hermansky, H., Kohn, P., Morgan, N., and Bayya, A. (**1992**). "RASTA-PLP speech analysis technique," in *Proceedings of ICASSP 1992*, Vol. 1, pp. 121–124.

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

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

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

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

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

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

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

Moritz, N., Anemuller, J., and Kollmeier, B. (**2011**). "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," in *Proceedings of ICASSP 2011*, pp. 5492–5495.

Moritz, N., Schädler, M. R., Adiloglu, K., Meyer, B. T., Jürgens, T., Gerkmann, T., Kollmeier, B., Doclo, S., and Goetze, S. (**2013**). "Noise robust distant automatic speech recognition utilizing NMF based source separation and auditory feature extraction," in *Proceeding of CHiME Workshop 2013*, Vancouver, British Columbia, Canada, pp. 1–6.

Nadeu, C., Macho, D., and Hernando, J. (**2001**). "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," Speech Commun. **34**, 93–114.

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

Schädler, M. R. (**2014**). “Reference Matlab implementations of feature extraction algorithms,” http://medi.uni-oldenburg.de/SGBFB (Last viewed January 14, 2015).

Schädler, M. R., Meyer, B. T., and Kollmeier, B. (**2012**). “Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition,” J. Acoust. Soc. Am. **131**, 4134–4151.

Schröder, J., Moritz, N., Schädler, M. R., Cauchi, B., Adiloglu, K., Anemüller, J., Doclo, S., Kollmeier, B., and Goetze, S. (**2013**). “On the use of spectro-temporal features for the IEEE AASP challenge ‘Detection and classification of acoustic scenes and events,’ ” in *Proceeding of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2013*, pp. 1–4.

Vertanen, K. (**2006**). “Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments,” Technical report, Cavendish Laboratory, University of Cambridge, Cambridge, UK.

Viikki, O., and Laurila, K. (**1998**). “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” Speech Commun. **25**, 133–147.

Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., and Matassoni, M. (**2013**). “The second ‘CHiME’ speech separation and recognition challenge: Datasets, tasks and baselines,” in *Proceedings of Workshop on Automatic Speech Recognition and Understanding (ASRU) 2013*, pp. 126–130.

Weide, R. L., and Rudnicky, A. (**2008**). “The CMU pronouncing dictionary,” available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict (Last viewed January 14, 2015).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., and Povey, D. (**2009**). “The HTK book” (for HTK version 3.4). Cambridge University Engineering Department, pp. 1–384.