Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition

Bernd T. Meyer[1,2], *Constantin Spille*[1], *Birger Kollmeier*[1], *Nelson Morgan*[2,3]

[1]Medical Physics, Carl-von-Ossietzky Universität Oldenburg, Germany
[2]International Computer Science Institute, Berkeley, CA, USA
[3]EECS Department, University of California - Berkeley, Berkeley, CA, USA

bernd.meyer@uni-oldenburg.de, constantin.spille@uni-oldenburg.de,
birger.kollmeier@uni-oldenburg.de, morgan@icsi.berkeley.edu

Abstract

Spectro-temporal filtering has been shown to result in features that can help to increase the robustness of automatic speech recognition (ASR) in the past. We replace the spectro-temporal representation used in previous work with spectrograms that incorporate knowledge about the signal processing of the human auditory system and which are derived from Power-Normalized Cepstral Coefficients (PNCCs). 2D-Gabor filters are applied to these spectrograms to extract features evaluated on a noisy digit recognition task. The filter bank is adapted to the new representation by optimizing the spectral modulation frequencies associated with each Gabor function. A comparison of optimized parameters and the spectral modulation of vowels shows a good match between optimized and expected range of frequencies. When processed with a non-linear neural net and combined with PNCCs, Gabor features decrease the error rate compared to the baseline and PNCCs by at least 19%.

Index Terms: automatic speech recognition, spectro-temporal features, power-normalized features

1. Introduction

The recognition of spoken language is a task that is easily performed by the healthy human auditory system, yet automatic speech recognition (ASR) often fails to accurately transcribe speech, especially in noisy conditions. It is this performance gap between human speech recognition (HSR) and ASR that motivates the use of auditory features for ASR with the aim of increasing its overall robustness.

Kleinschmidt and Gelbart [2] proposed the use of 2-dimensional Gabor filters to capture spectral, temporal and spectro-temporal cues from speech signals. This was motivated by psycho-acoustic and physiologic findings showing that these cues are detected by the primary auditory cortex of mammals, and are presumably also exploited by human listeners [4]. A challenge when designing filters for this task is to determine a set of suitable parameters that result in a robust feature set. Kleinschmidt and Gelbart used a machine learning approach that started with a random set of filters that was iteratively optimized using a simple classifier (a linear neural net) with a small speech recognition task (the recognition of isolated digits). In more recent work, Gabor filters have been organized in a filter bank (Figure 1), which resulted in relative improvements of the word error rate (WER) by 30-45% compared to a MFCC baseline for ASR [9, 5], and 21% for speaker identification [3]. However, one major drawback of this feature extraction scheme is that the filtering operates on mel-spectrogram, which is easily corrupted by noise. This representation incorporates some properties of the auditory system to a limited extent, but it also largely ignores physiological and psychoacoustic findings (such as the asymmetric filters for the place-frequency mapping of the basilar membrane in the inner ear).

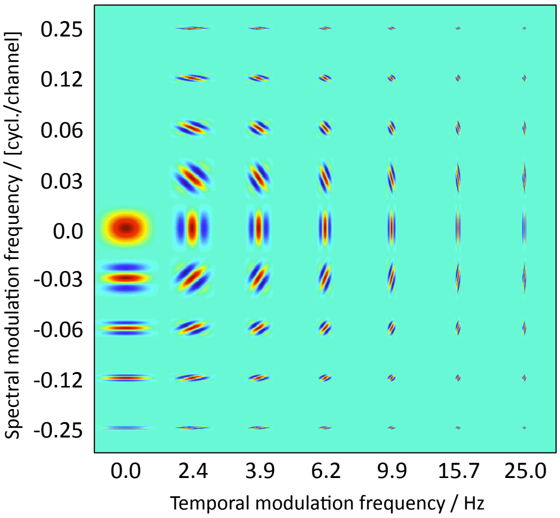
Spectral modulation frequency / [cycl./channel]
0.25
0.12
0.06
0.03
0.0
-0.03
-0.06
-0.12
-0.25
0.0
2.4
3.9
6.2
9.9
15.7
25.0
Temporal modulation frequency / Hz

Figure 1: Real components of Gabor filters used for the filter bank, arranged by temporal and spectral modulation frequencies.

In this paper, we replace the mel-spectrogram that has been used before [2, 9, 5, 3, 8] with a different representation borrowed from Power-Normalized Cepstral Coefficients (PNCCs), a feature type that has been shown to be quite robust against a large variety of noises [1], an example being shown in Figure 2. Replacing the spectro-temporal representation effectively changes the spectral modulation frequencies in terms of cycles/octave, which are associated with 2D Gabor filters. We therefore adapt the filters used in [5] by optimizing the distribution of these frequencies.

In the following, we first give an overview of power-normalized spectra and the feature calculation that is based on a Gabor filter bank. Next, the adaptation of modulation frequencies is described, as well as the ASR setup used to evaluate the spectro-temporal features. The optimized spectral modulation frequencies are compared to the modulations associated with the formant structure of vowel phonemes.

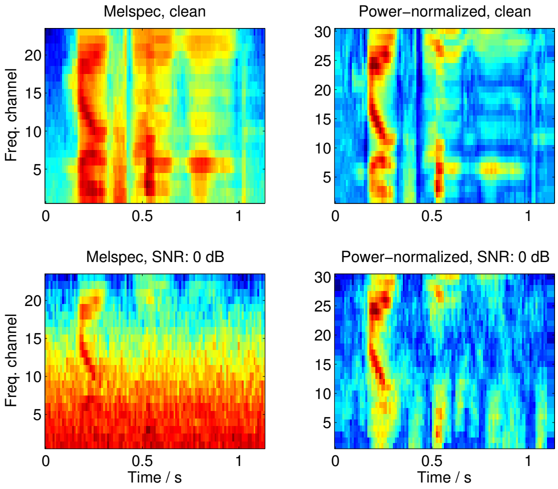

Figure 2: Clean and noisy mel-spectrogram and power-normalized spectrogram. In this example, speech-shaped, stationary noise has been added to the speech signal.

2. Feature extraction

2.1. Power-normalized spectrograms

Power-normalized spectrograms are obtained by performing the steps outlined in [1] to calculate Power-Normalized Cepstral Coefficients (PNCCs), but omitting the final discrete cosine transform and normalization.

First, a pre-emphasis filter is applied to a time signal and the STFT is calculated for 25 ms frames with a 10 ms frame shift. The magnitudes of the squared spectra are integrated via a Gammatone filter bank, which has been shown to better approximate the place-frequency mapping of the basilar membrane compared to the triangular filters used for MFCCs [7]. A power function nonlinearity that mimics the dependency of the input sound level and the perceived loudness is used to compress the output of the Gammatone filter bank. The exponent of 0.1 was derived from the relation of auditory-nerve firings and the level of a presented tone. Finally, the medium-time power bias is removed, which is motivated by the fact that the auditory system is sensitive to changes of the incoming signal (in contrast to low-modulated background noises, which are largely ignored). These enhancements result in a representation that is inherently more robust compared to mel-spectra (Figure 2).

The calculation of the medium-power over $2M + 1$ frames (cf. [1]) results in a feature output with $N - 2M$ time frames (N being the number of frames for common feature types, such as MFCCs. In this paper, we follow [1] and use $M = 2$). For combination with other feature types (e.g., in feature stream experiments), we prefer to pad the first and last frames of the spectrogram to obtain a representation with N frames.

2.2. Spectro-temporal features

Gabor features are calculated by processing a spectro-temporal representation of the input signal by a number of 2D modulation filters. Filtering is performed by calculating the 2D convolution of the filter and the power-normalized spectrogram.

Gabor filters are defined as the product of a complex sinusoidal function $s(n, k)$ (with n and k denoting the time and frequency index, respectively) and an envelope function $h(n, k)$. In this notation, the complex sinusoid is defined as

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

and the Hann function that we chose as envelope (with the parameters W_n and W_k for the window length) is given by

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

The periodicity of the carrier function is defined by the radian frequencies ω_k and ω_n, which allow the Gabor function to be tuned to particular directions of spectro-temporal modulation, including diagonal modulations. Experiments presented in this paper are based on a spectro-temporal filter bank proposed in [9]. The filter bank contains a set of temporal, spectral and spectro-temporal filters (Fig. 1) that were chosen to cover a wide range of modulation frequencies relevant in speech recognition. The result of the time-aligned convolution for all filters is used as feature vector. To limit the dimensionality of that vector, representative frequency channels are selected from the convolution, which is referred to as 'critical sampling' as described in [9, 5].

2.2.1. *Optimization of spectral modulation frequencies*

The parameters of the Gabor filter bank were optimized for mel-spectrograms as underlying spectro-temporal representation in [9], with spectral modulation frequencies being defined in cycles per channel. When this representation is replaced with power-normalized spectrograms, the center frequencies and hence the spectral modulation frequency of the Gabor filter bank in terms of cycles per *octave* are changed.

We adapt the filter bank to the new representation by optimizing the spectral distance d_k between filters. This parameter controls the distribution of spectral modulation frequencies ω_k which are given by

$$\omega_k^{i+1} = \omega_k^i \frac{1+c/2}{1-c/2}, \; c = 8d_k/\nu_k$$

(with ν_k being the number of semi-cycles of the carrier function under the envelope) [9]. The values for ω_k that result from setting d_k to 0.2 are depicted in Figure 1.

We analyze how the adapted filter frequencies relate to the spectral modulations associated with English vowels, compare this to the more intuitive unit cycles per octave (instead of cycl./ch.), and test if the optimization yields matched filters for the detection of specific vowels, as illustrated in Figure 3.

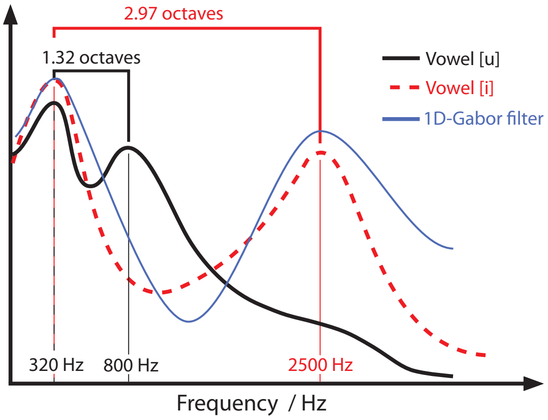

Figure 3: Schematic vowel spectra and illustration of 1D-Gabor filter.

2.3. ASR back-end

In [5], the MFCC baseline was improved when Gabor features were used as direct input to train and test a Hidden Markov Model (HMM). However, the lowest error rates were obtained when the features were non-linearly transformed with a multi-layer perceptron (MLP) and concatenated with MFCCs before HMM-based classification. In this work, we tested both approaches and present results for both recognition architectures (with PNCC features replacing MFCCs used before).

The performance of different realizations of Gabor features is tested within the Aurora 2 framework [6], which provides both speech data as well as specifications for the HMM classifier. It defines two training conditions that use either clean connected digits or a mixture of noisy and clean data (multi-condition) for training. Testing is performed with clean and noisy data using ten different noises. The average WERs are obtained by averaging the WERs of the test data with SNRs from 0 dB to 20 dB; the HMM was configured according to [6].

For experiments that employ non-linear weighting of features, the MLP training was carried out with phonetically labeled digit sequences from the Aurora 2 database. The phoneme labels were obtained from forced alignment. The MLP used 9 frames of temporal context which resulted in $9 \times N$ input units for N-dim. Gabor features. 160 and 56 units were used for the hidden and output layer, respectively. The log-posteriors were decorrelated with a principal component analysis in order to match the orthogonality assumption of the HMM decoder. Mean and variance were normalized for each utterance before training and testing the back end. For the last set of experiments, 13-dim. PNCC features with delta and acceleration coefficients were appended to the MLP-transformed Gabor features, resulting in 71-dim. feature vectors.

3. Results

3.1. ASR recognition scores

Results are compared to the MFCC baseline from Aurora 2, MFCCs and PNCCs with utterance-wise cepstral mean and variance normalization (CMVN), and the best feature set using Gabor features calculated from Mel-spectrograms presented in [5] (Table 1). Word error rates (WERs) for Gabor features based on power-normalized spectra are shown in Table 2. The variation of spectral modulation frequencies results in relatively small fluctuations of the WER. The optimal performance is reached for $d_k = 0.25$. For the combination with PNCCs, this optimum is shifted to $d_k = 0.5$, presumably because the modulation frequencies captured by these filters is complementary to the frequencies covered by PNCCs.

Error rates for Gabor features with MLP-processing (not shown in Table 2) were comparable to WERs obtained with Gabors fed directly to the HMM. However, when PNCCs are concatenated with the 32-dimensional MLP output, the WERs are further reduced and result in a relative improvement of up to 46% (multi train) and 71% (clean train) compared to the MFCC baseline. The improvements compared to the PNCC-based system are 26% and 19% for clean and multi-condition training.

3.2. Comparison with spectral modulations of vowels

To test if the filter optimization (Section 2.2.1) results in matched filters for vowels, spectral modulation frequencies (SMFs) of vowels are compared to SMFs of the Gabor filter bank (Table 3). The Gabor filters with the lowest

	Multi	Clean
MFCC	13.6	39.9
MFCC (CMVN)	9.5	20.7
PNCC	9.8	14.2
Gabor (IS11)	8.0	25.2

Table 1: Baseline word error rates for the training conditions 'Multi' and 'Clean'.

d_k	Gabor			MLP(Gabor) +PNCC (Dim: 71)	
	Dim.	Multi	Clean	Multi	Clean
0.20	787	8.3	14.5	7.6	12.1
0.22	761	8.4	14.3	7.5	11.9
0.25	644	**8.0**	**13.2**	7.5	12.0
0.30	631	8.1	14.6	7.4	11.6
0.40	527	8.1	14.8	7.5	11.6
0.50	462	8.3	14.4	**7.3**	**11.5**
0.60	436	8.3	14.3	7.4	12.0

Table 2: Word error rates for Gabor features based on power-normalized spectrograms. The varied parameter d_k controls the spectral spacing of filters. Results were obtained by using Gabors as direct input to an HMM or by additional MLP-processing and concatenation with PNCCs.

and highest SMF (4.29 and 25 cyc./ch. $\times 10^{-2}$) exhibit very similar SMFs to /i/ and /a/ and can therefore be interpreted as matched filters for the specific values listed in Table 3. The remaining filters do not correspond directly to the SMFs of other vowels (which might even be beneficial, since the presented vowel data is restricted to specific formant frequencies, while formant frequencies in speech signals exhibit a large variance). However, when comparing optimized SMFs with the vowel data, it becomes clear that the expected range of SMFs is well captured by spectral filters.

i) Formant frequencies, octave relation of F1/F2 and spectral MFs									
	i	e	y	ø	ɛ	u	o	ɑ	a
a) F_1	320	500	320	500	700	320	500	700	1000
b) F_2	2500	2300	1650	1500	1800	800	1000	1150	1400
c) $1/\omega_t$	2.97	2.20	2.37	1.59	1.36	1.32	1.00	0.72	0.49
d) $10^{-2}\omega_n$	4.55	5.88	5.88	8.33	9.09	11.1	14.3	16.7	25

ii) Spectral modulation freqs. of optimized filters (cycl./channel × 10^{-2})
4.29, 7.72, 13.89, 25

Table 3: The table presents frequencies for the first and second vowel formant in Hz (a and b), the number of octaves between F_1 and F_2 (c) and the corresponding spectral modulation frequency in cycl./channel$\times 10^{-2}$ (d). The latter can be compared to optimized spectral modulation frequencies ($d_k = 0.25$) of Gabor filters (ii).

4. Summary

In this paper, we presented ASR results obtained with spectro-temporal filters combined with power-normalized spectrograms, which are less sensitive to noise compared to mel-spectrograms used in previous work. The Gabor filter bank employed to extract spectro-temporal features was adapted to this new representation by optimizing the spectral modulation frequencies associated with filters. The optimized parameters were compared to the spectral modulation frequencies of vowel phonemes. While only some of the filters can be interpreted as matched filters for specific vowels, the range of frequencies fits well to the range observed for vowel formants. ASR experiments show that Gabor filters based on power-normalized spectrograms result in a robust feature set (especially when used in combination with neural nets and concatenation with PNCC features), with relative improvements of the WER by 19-26% compared to a PNCC baseline.

5. Acknowledgements

Supported by the DFG (SFB/TRR 31 'The active auditory system'; URL: http://www.uni-oldenburg.de/sfbtr31). We thank Suman Ravuri for supplying the MLP training scripts, and Marc René Schädler, Shuo-Yiin Chang, Chanwoo Kim, and Mark Harvilla for their support.

6. References

[1] Kim, C. Stern, R. M. (2009). "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in Proc. Interspeech, pp. 28-31.

[2] Kleinschmidt, M. and Gelbart, D. (2002). "Improving word accuracy with Gabor feature extraction," in Proc. Interspeech, pp. 25-28.

[3] Lei, H., Meyer, B., Mirghafori, N. (2012). "Spectro-temporal Gabor features for speaker recognition," in Proc. ICASSP.

[4] Mesgarani, N., Stephen, D., and Shamma, S. (2007). "Representation of phonemes in primary auditory cortex: how the brain analyzes speech," in Proc. ICASSP, pp. 765-768.

[5] Meyer, B. T., Ravuri, S. R., Schädler, M. R., Morgan, N. (2011). "Comparing different flavors of spectro-temporal features for ASR," in Proc. Interspeech, pp. 1269-1272.

[6] Hirsch, H. and Pearce, D. (2000). "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in Proc. ICSLP, Volume 4, pp. 29-37.

[7] Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M. (1992). "Complex sounds and auditory images," Auditory Physiology and Perception, Advances in Biosciences. Y. Cazals, L. Demany, K. Horner (eds.). Pergamon Press, Oxford, 1992, pp. 429443.

[8] Ravuri, S. and Morgan, N. (2010). "Using spectro-temporal features to improve AFE feature extraction for ASR," in Proc. Interspeech, pp. 1181-1184.

[9] Schädler, M.R., Meyer, B.T., Kollmeier. B. (2011). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," J. Acoust. Soc. Am., in press.