Comparing Different Flavors of Spectro-Temporal Features for ASR

Bernd T. Meyer[1], *Suman V. Ravuri*[1,2], *Marc René Schädler*[3], *Nelson Morgan*[1,2]

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

bmeyer@icsi.berkeley.edu, ravuri@icsi.berkeley.edu
marc.r.schaedler@uni-oldenburg.de, morgan@icsi.berkeley.edu

Abstract

In the last decade, several studies have shown that the robustness of ASR systems can be increased when 2D Gabor filters are used to extract specific modulation frequencies from the input pattern. This paper analyzes important design parameters for spectro-temporal features based on a Gabor filter bank: We perform experiments with filters that exhibit different phase sensitivity. Further, we analyze if non-linear weighting with a multi-layer perceptron (MLP) and a subsequent concatenation with mel-frequency cepstral coefficients (MFCCs) has beneficial effects. For the Aurora2 noisy digit recognition task, the use of phase sensitive filters improved the MFCC baseline, whereas using filters that neglect phase information did not. While MLP processing alone did not have a large effect on the overall performance, the best results were obtained for MLP-processed phase sensitive filters and added MFCCs, with relative error reductions of over 40% for both noisy and clean training.

Index Terms: spectro-temporal features, automatic speech recognition

1. Introduction

While human listeners are able to recognize speech even in very adverse acoustic conditions, automatic speech recognizers are much less robust in the presence of noise or channel distortions [5, 8]. It is this observation that motivates research to apply signal processing strategies of the human auditory system to automatic speech recognition (ASR). Physiological measurements in different mammal species have shown the existence of neurons in the primary auditory cortex (A1), which are sensitive to different patterns in the spectro-temporal representation of the signal [6]. The spectro-temporal receptive field (STRF) is an estimate of the stimulus that elicits a high firing rate in isolated neurons. A high percentage of STRFs exhibit patterns that span durations of 200 ms, which exceeds the time intervals considered by traditional ASR features. Furthermore, individual neurons were found to be sensitive to specific spectral and temporal modulation frequencies.

In physiological studies, two-dimensional Gabor functions have been successfully applied to modeling STRFs [10], which motivated the use of Gabor features as a front-end for ASR. Kleinschmidt and Gelbart [4] used a set of complex 2D Gabor functions for processing spectro-temporal representations of speech and found considerable improvements for ASR, especially for mismatched training and testing (i.e., using clean training data and noisy test utterances). This approach was also shown to be more robust against certain intrinsic variations of speech compared to standard features, such as changes in speaking rate [9]. In [4] and [9], filters were selected with the Feature Finding Neural Network that uses a simple classifier and a substitution rule which iteratively replaces the least relevant feature from a randomly drawn set. Filters were chosen from a feature space spanned by temporal and spectral modulation frequencies, the frequency channel, and the phase component of the feature (which is realized by selecting the imaginary, real part, or the magnitude of the complex filter result as shown for a 1D example in Fig. 1). The optimized set was used as input to a Tandem recognition system, which combines a multilayer perceptron (MLP) and a Hidden Markov model (HMM) as back end [2]. A challenge often encountered in feature selection is that this approach may result in feature sets that are optimized for a specific task, while the dynamic weighting of a large number of spectro-temporal features might be more appropriate when a feature set is to be used in different acoustic backgrounds. In [11], a multitude of filters ($>$ 10k) was used to extract spectro-temporal modulation features that are organized in streams; these streams were processed with multilayer perceptrons (MLPs), and individual streams were merged and combined with standard features. In another approach, a Gabor filter bank was used to extract spectro-temporal features that were reported to improve an MFCC baseline [1] when the real-valued filter output was used as direct input for an HMM classifier [12]. These studies share the idea of using using localized filters tuned to specific spectro-temporal pattern (e.g., vowel transients or changes of the fundamental frequency as

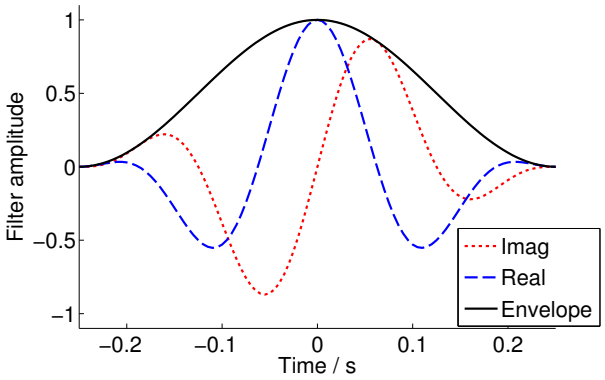

Figure 1: Real and imaginary part of a 1-dim. Gabor function. While the zero phase of the real component preserves the location of matched patterns in the filter output, the imaginary component can serve as an edge detector for the signal analyzed.

shown in the example in Fig. 2), thereby increasing the overall robustness of an ASR system. However, the studies also differ with respect to the design of spectro-temporal filters, and it is unclear how these affect the overall performance obtained from the derived feature sets. One of the design choices is the relevance of phase information, which was subject to the feature selection process in [4], whereas other works use a fixed phase for all filters ([11], [12]). A second aspect is the non-linear weighting of features, which was applied in [4] and [11], while in another case the Gabor features were used to directly train and test a Hidden Markov Model [12]. Finally, a combination of Gabor and standard features was proposed in [11], which improved a competitive baseline system for a digit recognition task.

This work presents a series of experiments in which these different design choices are investigated (cf. Fig. 3). The starting point for ASR tests is a filter bank proposed in [12] that evenly covers the modulation frequency domain, which avoids the problem of selecting filters for a specific recognition task. We analyze the importance of phase sensitivity of spectro-temporal filters by using the imaginary or real part or the magnitude of the filter output. The results are compared to non-linearly weighted feature streams that are calculated using MLPs. The MLPs produce phone posterior probabilities, which are combined with mel-frequency cepstral coefficients (MFCCs) in a final set of experiments.

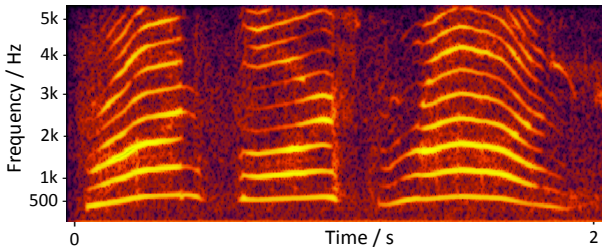

Figure 2: Spectrogram of the utterance "Run Forrest, run", which exhibits spectro-temporal modulations that arise from vowel transients and variations of the fundamental frequency.

2. Methods

2.1. Gabor 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 a log mel spectrogram; the latter was chosen because it incorporates several properties of the auditory system (i.e., non-linear frequency scaling and logarithmic compression of amplitude values).

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 Gaussian envelope function. In this work, the Gaussian envelope is replaced by the Hann function $h(n, k)$, which was reported earlier to slightly improve results for ASR [7] due to improved filter characteristics compared to a Gaussian envelope with limited extent. In this notation, the complex sinusoid is defined as

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

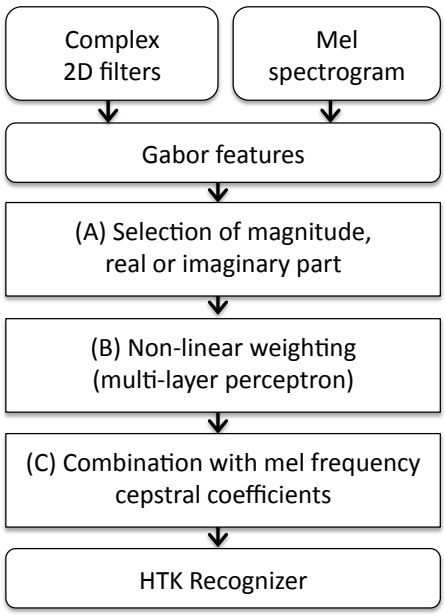

Figure 3: Diagram of processing steps for ASR with complex spectro-temporal features. This paper analyzes the effect of choosing filters with different phase sensitivity (A), non-linear weighting of features (B), and combination with standard features (C).

and the Hann envelope is given by

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

and the sinusoidal function $s(n, k)$ with the window lengths W_n and W_k.

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. For purely temporal or spectral filters, this definition results in an infinite support function; in these cases, the support is limited to 69 frequency channels or 99 time frames, which corresponds to the maximum size of the other filters in the respective dimension.

2.2. Gabor filter bank

Experiments presented in this paper are based on a spectro-temporal filter bank proposed in [12]. The filter bank contains a set of temporal, spectral and spectro-temporal filters that were chosen to cover a wide range of modulation frequencies. The specific modulation frequencies were chosen so that the transfer functions of the filters exhibit a constant overlap in the modulation frequency domain; these frequencies and center frequencies of the mel spectrograms are listed in Table 1. While the lowest temporal modulation frequency employed in [12] was 6 Hz, we use a modified version in which the lowest modulation frequency is 2 Hz, which was included to cover modulations arising from the syllable structure in spoken language. This parametrization results in 59 pairs of spectral and temporal modulation frequencies; the resulting filters are depicted in Fig. 4.

With 59 spectro-temporal filters and 23 frequency channels, the resulting feature vectors have 1357 components, which is

Temp. mod [Hz]	0, 1.9, 3.9, 6.2, 9.9, 15.7, 25
Spec. mod. [cycl./oct.]	-0.25, -0.1224, -0.06, -0.0293, 0, 0.0293, 0.06, 0.1224, 0.25
Center freq. [Hz]	124, 189, 260, 336, 417, 506, 601, 704, 814, 934, 1063, 1202, 1352, 1515, 1689, 1878, 2082, 2302, 2539, 2794, 3070, 3368, 3689, 4036, 4410, 4814, 5249, 5719, 6226, 6773, 7363

Table 1: Temporal and spectral modulation frequencies used for the filter bank and center frequencies of the mel spectrograms used as spectro-temporal representation for feature calculation.

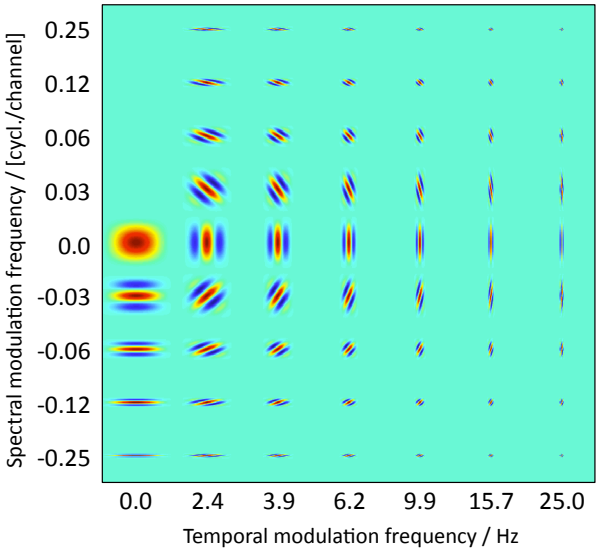

Figure 4: Real components of Gabor filters used for the filter bank, arranged by temporal and spectral modulation frequencies.

too high-dimensional to be used with the Aurora2 HMM. However, since filters with a large spectral extent result in relatively small changes in the feature values when shifted by one frequency channel, the redundancy of the filter output can be reduced by selection of specific feature channels. Hence, for each modulation filter, the center frequency channel (corresponding to a frequency of 1 kHz) is selected; additionally, the channels are included in the final vector for which the overlap of neighboring Gabor filters is 3/4. With this critical sampling, the number of selected channels lies between 1 (for ω_k = 0 cycl./oct.) or 23 (ω_k = $\pm$ 0.25 cycl./oct.), and the feature dimension is reduced to 449.

2.3. Classifier and baseline

The performance of different realizations of Gabor features is tested within the Aurora2 framework [3], which provides both speech data as well as specifications for the HMM classifier. Aurora2 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, with a mixture of various noisy types: Subway, babble, car, exhibition (which are also used during multi-condition training), restaurant, street, airport, and station. The average word error rates (WERs) reported for this task are obtained by averaging the WERs of the test data with SNRs from 0 dB to 20 dB. Additionally, the relative improvements in WER over the baseline system are presented. These were obtained using MFCCs with delta and acceleration coefficients as input features. The HMM was configured according to [3]: The setup uses whole-word HMMs with 16 states and with a 3-Gaussian mixture with diagonal covariances per state. Skips over states are not permitted in this model.

For experiments that employ non-linear weighting of features, the MLP training was carried out with phonetically labeled digit sequences from the Aurora2 database. The phoneme labels were obtained from forced alignment. The MLP used 9 frames of temporal context which resulted in $9 \times 449 = 4041$ input units. 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-dimensional MFCC features with delta and acceleration coefficients were appended to the MLP-transformed Gabor features, resulting in 71-dimensional feature vectors.

3. Results

The results of the ASR experiments are presented in Table 2. All implementations of Gabor features decreased the baseline error rate for both training conditions except for the case when filters that are insensitive to the phase were used as direct input to an HMM back end. When using MLP-weighted features, error rates went slightly up in three out of six cases. For the magnitude filter output however, error rates were lowered by 8% and 20% absolute compared to Gabor features without non-linear weighting. With MLP processing, the large differences between real, imaginary part and magnitude were not observed. Errors were further reduced when the phone posteriors from the MLP were concatenated with MFCCs, with error rates of 25.2% and 8.0% for real-valued phase-sensitive filters. For consistency, relative improvements were calculated from the WERs in this table. When the measure proposed in [3] is used (which takes the each individual WER for the noise types and SNRs into account), identical trends are observed with the exception of 'Gabor+MLP+MFCC (imag)', which resulted in a higher relative improvement (55%) compared to 'Gabor+MLP+MFCC (real)' (49%).

4. Discussion and summary

The ASR results show that phase information is an important design aspect when spectro-temporal features are used as direct input to an HMM classifier. For the filter bank used in this work, the use of phase-sensitive filters outperformed using the magnitude by 8.8% and 16% absolute for clean and multi-condition training, respectively. While the imaginary component might be able to serve as edge detector in the spectro-temporal domain, the real component is designed to capture spectro-temporal modulations in any possible direction - including simple temporal or spectral modulations. Since it is sensitive to the phase, it is a good estimate for the temporal and spectral location of events. The enhanced localization with

	Absolute WER		Rel. imp. in WER	
	Clean	Multi	Clean	Multi
MFCC Baseline	42.7	15.6	0.0	0.0
Gabor (Real)	30.5	11.2	28.6	28.2
Gabor (Imag)	36.0	11.7	15.7	25.0
Gabor (Mag)	52.8	20.0	-23.7	-28.2
+MLP (Real)	30.7	12.6	28.1	19.0
+MLP (Imag)	31.4	13.2	26.5	15.6
+MLP (Mag)	32.6	12.0	23.7	23.4
+MLP+MFCC (Real)	25.2	8.0	41.1	48.4
+MLP+MFCC (Imag)	27.2	8.2	36.3	47.5
+MLP+MFCC (Mag)	30.7	8.3	28.1	47.0

Table 2: ASR word error rates (WER) and relative improvement of WER for different implementations of Gabor features. 'Clean' and 'Multi' refer to the training conditions for Aurora2 digit recognition task.

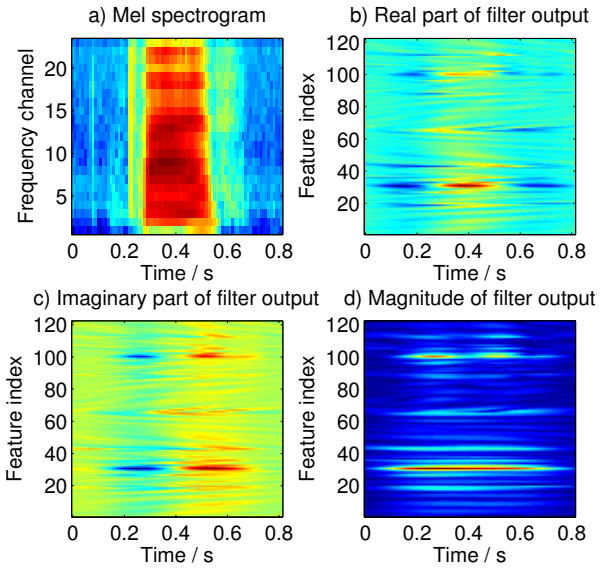

Figure 5: Mel spectrogram of the utterance five (a). Gabor filter bank features for real-valued or imaginary filter output (b and c) and for the magnitude of the filter result (d).

real-valued filters in comparison to other variations of Gabor features in depicted in Fig. 5. In this figure, filters with temporal modulations below 6 Hz were selected, for which the dislocation of energy is clearly visible for the imaginary component and the magnitude. This result suggests spending more attention to the phase component of spectro-temporal filters in future experiments. It is likely that information from streams that exhibit different phase sensitivity is complementary, which could be exploited by merging these streams with neural networks. In this scenario, the addition of a free phase parameter may also be considered (in contrast to using filters with a phase of 0 or $\pi/2$ as in this study).

The large differences between real, imaginary part and the magnitude were strongly decreased when Gabor features were nonlinearly transformed with an MLP (or in one case even inverted, cf. Table 2). It might be that the temporal context used for the MLP to some extent counteracts the dislocation of energy. The addition of MFCCs features lowered the WER by 4.2% on average, which is consistent with other studies [11]. This result also suggests that the error rates for Gabor features in combination with MLPs or other techniques for feature weighting might be further decreased since the filter bank generating the feature also include purely spectral filters, which resemble the processing performed by MFCCs.

5. Acknowledgements

Bernd T. Meyer's work is supported by a post-doctoral fellowship of the German Academic Exchange Service (DAAD). We also thank the National Defense Science and Engineering Graduate Fellowship (NDSEG) for helping to fund Suman Ravuri's research, and Cisco Systems, Inc. for funding Nelson Morgan's work.

6. References

[1] Davis, S. and Mermelstein, P. (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, 28 (4), pp. 357-366.

[2] Hermansky, H., Ellis, D., and Sharma, S. (2000). "Tandem connectionist feature extraction for conventional HMM systems," in Proc. Interspeech, pp. 1635-1638.

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

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

[5] Lippmann, R. (1997). "Speech recognition by machines and humans," Speech Commun. 22 (1), 1-15.

[6] 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.

[7] Meyer, B. T. and Kollmeier, B. (2008). "Optimization and Evaluation of Gabor feature sets for ASR," in Proc. Interspeech, pp. 906-909.

[8] 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, pp. 388-403.

[9] Meyer, B.T. and Kollmeier, B. (2011). "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," Speech Communication 53, pp. 753-767.

[10] Qiu, A., Schreiner, C., and Escabi, M. (2003). "Gabor analysis of auditory mid- brain receptive fields: spectro-temporal and binaural composition," Journal of Neurophysiology, 90, pp. 456-476.

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

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