Spectro-temporal Gabor features as a front end for automatic speech recognition

Pacs reference 43.72

Michael Kleinschmidt
Universität Oldenburg
- Medizinische Physik -
D-26111 Oldenburg
Germany
Phone: ++49 441 798 3146
Fax : ++49 441 798 3902
Email: michael@medi.physik.uni-oldenburg.de

International Computer Science Institute
1947 Center Street
Berkeley, CA 94704
USA

ABSTRACT

A novel type of feature extraction is introduced to be used as a front end for automatic speech recognition (ASR). Two-dimensional Gabor filter functions are applied to a spectro-temporal representation formed by columns of primary feature vectors. The filter shape is motivated by recent findings in neurophysiology and psychoacoustics which revealed sensitivity towards complex spectro-temporal modulation patterns. Supervised data-driven parameter selection yields qualitatively different feature sets depending on the corpus and the target labels. ASR experiments on the Aurora dataset show the benefit of the proposed Gabor features, especially in combination with other feature streams.

INTRODUCTION

ASR technology has seen many advances in recent years, still the issue of robustness in adverse conditions remains largely unsolved. Additive noise as well as convolutive noise in the form of reverberation and channel distortions occur in most natural situations, limiting the feasibility of ASR systems in real world applications. Standard front ends, such as mel cepstra or perceptual linear prediction, only represent the spectrum within short analysis frames and thereby neglect very important dynamic patterns in the speech signal. This deficiency has been partly overcome by adding temporal derivatives in the form of delta and delta-delta features to the set. In addition, channel effects can be reduced by carrying out further temporal bandpass filtering such as cepstral mean substraction or RASTA processing [Her94]. A completely new school of thought has been initiated by a review of Fletcher's work [All94], who found log sub-band classification error probability to be additive for nonsense syllable recognition tasks observed on human subjects. This suggests independent processing in a number of articulatory bands without recombination until a very late stage. The most extreme example of the new type of purely temporal features are the TRAPS [Her98] which apply multi-layer perceptrons (MLP) to classify current phonemes in each single critical band based on a temporal context of up to 1s. Another approach is multi-band processing [Bou96], for which features are calculated in broader sub-bands to reduce the effect of band-limited noise on the overall performance. All these feature extraction methods apply *either* spectral or temporal processing at a time. Nevertheless, speech and many other natural sound sources exhibit distinct spectro-temporal amplitude modulations (see Fig. 2a as an example). While the temporal modulations are mainly due to the syllabic structure of speech, resulting in a bandpass characteristic with a peak around 4Hz,

spectral modulations describe the harmonic and formant structure of speech. The latter are not at all stationary over time. Coarticulation and prosody result in variations of fundamental and formant frequencies even within a single phoneme. This raises the question whether there is relevant information in amplitude variations oblique to the spectral and temporal axes and how it may be utilized to improve the performance of automatic classifiers. In addition, recent experiments about speech intelligibility showed synergetic effects of distant spectral channels [Gre98] that exceed the log error additivity mentioned earlier and therefore suggest spectro-temporal integration of information. This is supported by a number of physiological experiments on different mammal species which have revealed the spectro-temporal receptive fields (STRF) of neurons in the primary auditory cortex. Individual neurons are sensitive to specific spectro-temporal patterns in the incoming sound signal. The results were obtained using reverse correlation techniques with complex spectro-temporal stimuli such as checkerboard noise [deC98] or moving ripples [Sch00, Dep01]. The STRF often clearly exceed one critical band in frequency, have multiple peaks and also show tuning to temporal modulation. In many cases the neurons are sensitive to the direction of spectro-temporal patterns (e.g. upward or downward moving ripples), which indicates a combined spectro-temporal processing rather than consecutive stages of spectral and temporal filtering. These findings fit well to psychoacoustical evidence of early auditory features [Kae00], yielding patterns that are distributed in time and frequency and in some cases comprised of several unconnected parts. These STRF can be approximated, although somewhat simplified, by two-dimensional Gabor functions, which are localized sinusoids known from receptive fields of neurons in the visual cortex [deV90].

In this paper, new two-dimensional features are investigated, which can be obtained by filtering a spectro-temporal representation of the input signal with Gabor-shaped localized spectro-temporal modulation filters. These new features in some sense incorporate but surely extend the features mentioned above. A recent study showed an increase in robustness when real valued Gabor filters are used in combination with a simple linear classifier on isolated word recognition tasks [Kle02]. Now, the Gabor features are modified to a complex filter and based on mel-spectra, which is the standard first processing stage for most types of features mentioned above. It is investigated whether the use of Gabor features may increase the performance of more sophisticated state-of-the-art ASR systems. The problem of finding a suitable set of Gabor features for a given task is addressed and optimal feature sets for a number of different criteria are analyzed.

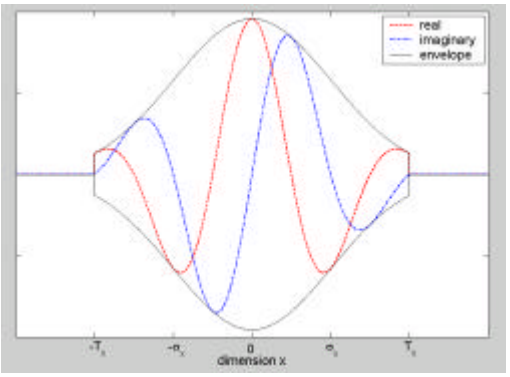


Figure 1:
Example of a one-dimensional complex Gabor function or a cross section of a two-dimensional one. Real and imaginary components are plotted, corresponding to zero and ð/2 phase, respectively. Note, that one period $T_x=2ð/ù_x$ of the oscillation fits into the interval [-ó$_x$ ó$_x$] and the support in this case is reduced from infinity to twice that range or $2T_x$. An example of a 2D-Gabor function can be found in Fig. 2b.

GABOR FILTER FUNCTIONS

The Gabor approach pursued in this paper has the advantage of a neurobiological motivated prototype with only few parameters which allows for efficient automated feature selection. The parameter space is wide enough to cover a large variety of cases: purely spectral features are identical to sub-band cepstra - modulo the windowing function - and purely temporal features closely resemble the TRAPS pattern or the RASTA impuls response and its derivatives [Her98b]. Gabor features are derived from a two-dimensional input pattern, typically a series of feature vectors. A number of processing schemes may be considered for these primary features that extract a spectro-temporal representation from the input wave form. The range is from a spectrogram to sophisticated auditory models. In this study the focus is on the log mel-spectrogram for its widespread use in ASR, and because it can be regarded as a very simple

auditory model, with instantanious logarithmic compression and mel-frequency axis. In this paper, the log mel-spectrum was calculated as in [ETS00]. The processing consists of DC removal, Hanning windowing with 10ms offset and 25ms length, pre-emphasis, FFT and summation of the magnitude values into 23 mel-frequency channels with center frequencies from 124 to 3657Hz. The amplitude values are then compressed by the natural logarithm. The receptive field of cortical neurons is modeled by two-dimensional complex Gabor functions *g(t,f)* defined as the product of a Gaussian envelope *n(t,f)* and the complex Euler function *e(t,f).* The envelope width is defined by standard deviation values σ_f and σ_t, while the periodicity is defined by the radian frequencies ω_f and ω_t with *f* and *t* denoting the frequency and time axis, respectively. Further parameters are the centers of mass of the envelope in time and frequency t_0 and f_0. In this notation the Gabor function *g(t,f)* is defined as

$$g(t,f) = \frac{1}{2\pi\sigma_f\sigma_t} \cdot \exp\left(\frac{-(f-f_0)^2}{2\sigma_f^2} + \frac{-(t-t_0)^2}{2\sigma_t^2}\right) \cdot \exp\left(i w_f (f-f_0) + i w_t (t-t_0)\right)$$

It is reasonable to set the envelope width depending on the modulation frequencies in order to keep the same number of periods in the filter function for all frequencies. Basically, this makes the Gabor feature a wavelet prototype with a scale factor for each of the two dimensions. The spread of the Gaussian envelope in dimension *x* was set to $\sigma_x = \pi/\omega_x = T_x/2$ to have a full period T_x in the range between $-\sigma_x$ and σ_x as depicted in Fig. 1. The infinite support of the Gaussian envelope is cut off at σ_x to $2\sigma_x$ from the center. For time dependent features, t_0 is set to the current frame, so three main free parameters remain: f_0, ω_f and ω_t. The range of parameters is limited mainly by the resolution of the primary input matrix (100Hz and 23 channels covering 7 octaves). The temporal modulation frequencies were limited to a range of 2-50Hz, and the spectral modulation frequencies to a range of 0.04-0.5 cycles per channel or approximately 0.14-1.64 cycles per octave. If ω_f or ω_t is set to zero to obtain purely temporal or spectral filters, respectively, σ_t or σ_f again becomes a free parameter.

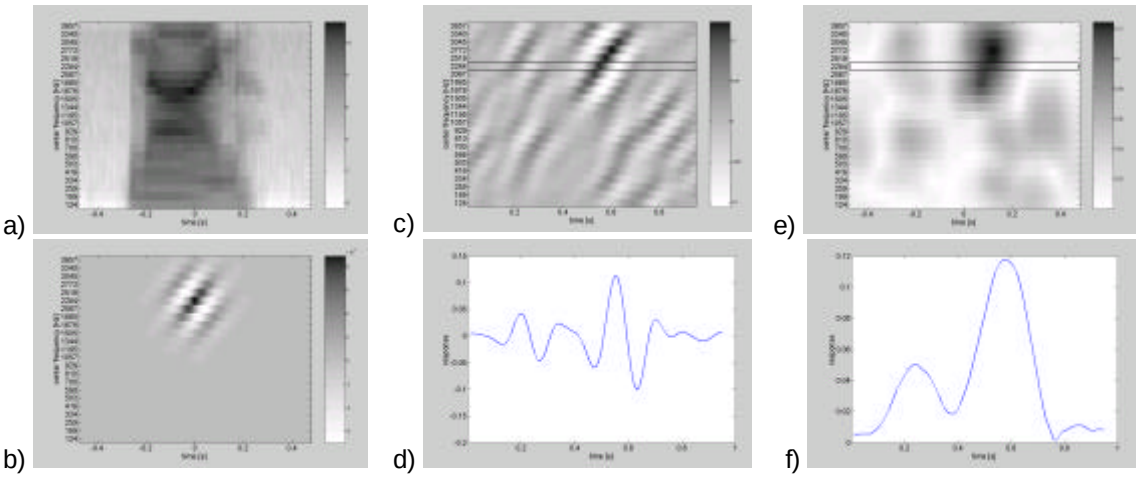


Figure 2: **a)** mel-scale log magnitude spectrogram of a *"Nine" from the TIDigits corpus.* **b)** *an example of a 2D-Gabor complex filter function (real values plotted here) with parameters –7Hz and 0.2 cycl./channel. The resulting filtered spectrograms for* **c)** *real and* **e)** *complex valued filters.* **e)** and **f)**: *The resulting feature values for f_0=2284Hz.*

From the complex results of the filter operation, real valued features may be obtained by using the real or imaginary part only. This method was used in [Kle02] and offers the advantage of being sensitive to the phase of the filter output and thereby to exact temporal location events. Alternatively, the magnitude of the complex filter output may be used. This gives a more smooth filter response (cf. Fig. 2f) and allows for a phase independent feature extraction which might be advantageous in some cases. Both type of filters have been used in the experiments below. The filtering is performed by calculating the correlation function over time of each input frequency channel with the corresponding part of the Gabor function and a subsequent summation over frequency. This yields one output value per frame per Gabor filter and is equivalent to a two-dimensional correlation of the input representation with the complete filter function and a subsequent selection of the desired frequency channel f_0 (see Fig.2).

FEATURE SELECTION

Due to the large number of possible parameter combinations, it is necessary to select a suitable set of features. This was carried out by a modified version of the Feature-finding neural network (FFNN). It consists of a linear single-layer perceptron in conjunction with secondary feature extraction and an optimization rule for the feature set [Gra90]. The linear classifier guarantees fast training, which is necessary because in this wrapper method for feature selection the importance of each feature is evaluated by the increase of RMS classification error after its removal from the set. This 'substitution rule' method [Gra91] requires iterative re-training of the classifier and replacing the least relevant feature in the set with a randomly drawn new one. When using the linear network for digit classification without frame by frame target labeling temporal integration of features is necessary. This is done by simple summation of the feature vectors over the whole utterance yielding one feature vector per utterance as required for the linear net. The FFNN approach has been successfully applied to isolated digit recognition with the sigma-pi type of secondary features [Gra90] and also in combination with Gabor features [Kle02].

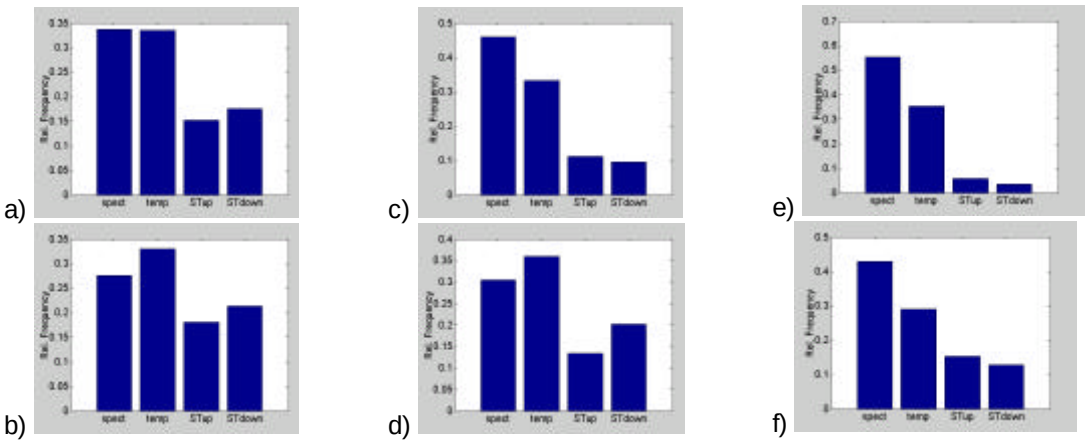
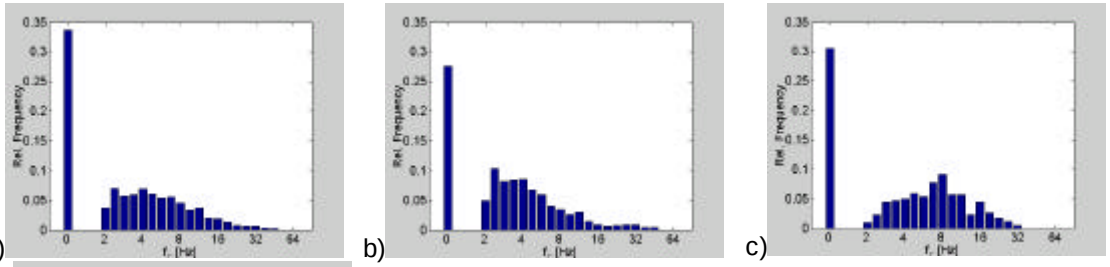


Figure 3:
Distribution of Gabor types **a)** *in all selected sets (103 sets with 2702 features) and* **b)** *for digits (43/1440),* **c)** *phone (38/836) and* **d)** *diphone (22/426) targets only. Overall percentages of spectral, temporal and spectro-temporal (ST) features are given. 'down' denotes negative temporal modulation. Distribution of Gabor types for phone targets with grouping into* **e)** *broad phonetic (manner) classes (8/152) and* **f)** *for single phonetic classes (18/476).*

Optimization was carried out on German and English digits targets (zifkom and TIDigits corpora), which are comprised of mainly monosyllabic words, as well as on parts of the TIMIT corpus with phone-based labeling on a frame by frame basis. The phone labels were grouped into a smaller number of classes based on different phonetic features (place and manner of articulation) or, alternatively, only members of a certain single phonetic class (e.g. vowels) were used in the optimization. In addition, optimization experiments were carried out with diphone targets, focusing on the transient elements by using only a context of 30ms to each side of the phoneme boundary. Again, target labels were combined to make the experiments feasible. More than 100 optimization runs were carried out on different data and with different target sets, each resulting in an optimized set of between 10 and 80 features. Apart from the free parameters f_0, $\grave{u}_f$ and $\grave{u}_t$ the filter mode (real, imaginary or complex) and filter type (spectral only, temporal only, spectro-temporal up, spectro-temporal down) were also varied and equally likely when randomly drawing a new feature.

The complex filter function (47.7% of all selected features) was consistently preferred over using the real or imaginary part only. This trend is most dominant for ST or purely temporal features, while for spectral features all modes are equally frequent. As can be seen in Fig. 3a, spectro-temporal (ST) features were selected in 32.7% of all cases. Only minor differences are found in average between using clean or noisy data for the optimization, but significant differences can be observed depending on the classification targets. ST features account for 39% of all features in the selected sets for digit target, while the numbers for diphone and phone targets are 33% and 21%, respectively.

a) b) c)



b)

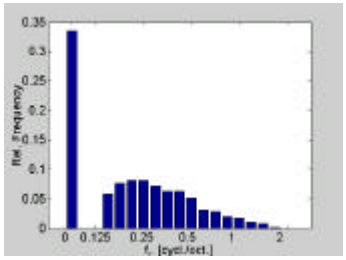


Figure 4:
Distribution of temporal modulation frequency over all Gabor types **a)** *in all selected sets,* **b)** *for digits and* **c)** *for diphone targets. Purely spectral features accumulate in the 0Hz bin, although they also have a limited temporal extend.* **d)** *Distribution of spectral modulation frequency for all targets. Purely temporal features accumulate in the 0 bin, although they also have a limited spectral extend.*

There is a significant difference between the phone targets which are grouped according to manner of articulation with necessary intergroup discrimination and those where only targets of one phonetic class were to be classified. In the former case, ST features were almost never selected (9%), while in the latter 28% of all features were ST, with highest number for diphthongs (46%) and lowest for stops (14%). For vowels spectral features dominated (56%) while for stops and nasals the percentage of temporal Gabor functions was highest (41% in both cases). The feature distribution along the parameter axis of temporal and spectral modulation are plotted in Fig. 4 a) and b). Please note that the parameter values were drawn from a uniform distribution over the log2 of the modulation frequencies. Temporal modulation frequencies between 2-8Hz dominate with lower modulation frequencies preferred for digit targets and medium (around 8Hz) for diphone targets. Spectral modulation frequencies are consistently preferred to be in the region of 0.2 to 0.7 cycles per octave with only minor differences across target labels. These results correspond well with the importance of different modulation frequencies for speech recognition [Kan98], modulation perception thresholds [Chi99] and physiological data [Mil02].

System description	WER [%]		WER red. [%]	
	multi	clean	multi	clean
R0: Aurora2 reference	12.97	41.94	0.00	0.00
R1: Melspec Tandem	12.04	28.66	12.87	40.09
G1: Gabor phone optimized	11.68	30.17	14.52	37.19
G2: Gabor digit optimized	11.99	23.63	4.03	51.24
RD:concatenate R1 & melspec diphone	12.86	32.48	8.97	32.38
G1D: concatenate G1 & Gabor diphone	11.17	25.29	19.74	50.57
RP: post. combination R1 + mel cepstra	13.45	33.20	13.91	45.08
G1P: post. combination G1+ R1	10.74	24.78	24.64	51.88
G2P: post combination G2 + R1	10.62	24.73	23.11	53.06
RQ: concatenate R0 & R1	10.74	29.06	25.50	41.98
G1Q: concatenate R0 & G1	10.35	27.89	30.45	48.39

Gabor set G1 was optimized on noisy TIMIT with broad phonetic classes, G2 on noisy German digits (zifkom).

Table 1: *Word error rate (WER) in percent and WER reduction relative to the Aurora2 baseline features (***R0***). WER and WER reduction are averaged separately over all test conditions. Non-Gabor reference system have gray shading.* **P** *denotes posterior combination of two Tandem streams before the final PCA.* **D** *indicates the concatenation of two Tandem streams which are optimized on phone and diphone targets, respectively, after reducing the dimension of each to 30 via PCA.* **Q** *indicates concatenation of R0 (42 mfcc features) with 18 Tandem features.* **R1** *denotes the Tandem reference system with MLP trained on mel-spectra features in 90ms of context.*

ASR EXPERIMENTS

Recognition experiments were carried out within the Aurora2 experimental framework (see [Hir00] for details). A fixed HTK back end was trained on multicondition (4 types of noise, 5 SNR levels) or clean only training data. Strings of English digits (from the TIDigits corpus) were then recognized in 50 different noise conditions with 1000 utterances each (10 types of noise and SNR of 0, 5, 10, 15, 20) including convolutional noise. The Tandem recognition system [Her00]

was used for the Gabor feature sets. Every set of 60 Gabor features is online normalized and combined with delta and double-delta derivatives before feeding into the MLP (60, 1000 and 56 neurons in input, hidden and output layer, respectively), which was trained on the TIMIT phone-labeled database with artificially added noise. The 56 output values are then decorrelated via PCA (statistics derived on clean TIMIT) and fed into the HTK back end.

The results in Tab. 1 show a drastic improvement of performance over the reference system (R0) by using the Tandem system, which is further increased by applying Gabor feature extraction (G1, G2) instead of mel-spectra (R1) or mel-cepstra (not shown). Even better performance is obtained by combining Gabor feature streams with mel-spectrum based feature streams via posterior combination (G1P, G2P, [Ell00]). Alternatively, improvement may be obtained by concatenation of a Gabor stream with another, diphone-based Gabor stream (G1D) or with the reference stream (G1Q). In all cases the combination of a Gabor feature stream with a non-Gabor stream yields better performance than combining two non-Gabor streams.

SUMMARY

An efficient method of feature selection is applied to optimize a set of Gabor filter functions. The underlying distribution of importance of spectral and temporal modulation frequency reflects the properties of speech and is in accordance with physiological and psychoacoustical data. The optimized sets increase the robustness of the Tandem digit recognition system on the TIDigits corpus. This is especially true when several streams are combined by posterior combination or concatenation, which indicates that the new Gabor features carry complementary information to that of standard front ends.

A major part of this work was carried out at the International Computer Science Institute in Berkeley, California. Special thanks go to Nelson Morgan, Birger Kollmeier, Steven Greenberg, Hynek Hermansky, David Gelbart, Barry Yue Chen, and Stephane Dupont for their support and many enlightening discussions. This work was supported by Deutsche Forschungsgemeinschaft (KO 942/15).

BIBLIOGRAPHY

[All94] J.B. Allen: "How Do Humans Process and Recognize Speech", *IEEE Trans. SAP* 2(4) 1994 pp. 567-576.
[Bou96] H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan, "Towards sub-band-based speech recognition," *European Signal Proc. Conf., Trieste*, 1996, pp. 1579–1582.
[deC98] R. C. deCharms, D. T. Blake, and M. M. Merzenich, "Optimizing sound features for cortical neurons," *Science*, vol. 280, pp. 1439–1443, 1998.
[Dep01] D.A. Depireux, J.Z. Simon, D.J. Klein, and S.A. Shamma: "Spectro-Temporal Response Field Characterization With Dynamic Ripples in Ferret Primary Auditory Cortex" *J. Neurophysiol.* 85, pp. 1220-1234, 2001.
[deV90] R De-Valois and K. De-Valois: "Spatial Vison", Oxford U.P., New York, 1990.
[Ell00] D.P.W. Ellis "Improved recognition by combining different features and different systems", *AVIOS 2000*.
[ETS00] Standard: ETSI ES 201 108 V1.1.2 (2000-04)
[Gra90] T. Gramß and H. W. Strube, "Recognition of isolated words based on psychoacoustics and neurobiology," *Speech Communication* 9, pp. 35–40, 1990.
[Gre98] S. Greenberg, T. Arai, and R. Silipo: "Speech intelligibility derived from exceedingly sparse spectral information*", Proc. ICSLP* 1998.
[Her94] H. Hermansky and N. Morgan: "RASTA processing of speech", *IEEE Trans. SAP* 2(4) 1994 pp. 578-589.
[Her98a] H. Hermansky and S. Sharma: "TRAPS - Classifiers of temporal patterns," *Proc. ICSLP'98*, 1998, vol. 3, pp. 1003–1006.
[Her98b] H. Hermansky: "Should recognizers have ears?", *Speech Communication* 25, pp. 3-27, 1998.
[Her00] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," *Proc. ICASSP 2000*, Instanbul, 2000.
[Hir00] H.G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions*," ISCA ITRW ASR2000, Paris - Automatic Speech Recognition: Challenges for the Next Millennium, 2000.*
[Kae00] C. Kaernbach, "Early auditory feature coding", *Contributions to psychological acoustics: Results of the8th Oldenburg Symposium on Psychological Acoustics*. 2000, pp. 295–307, BIS, Universität Oldenburg.
[Kan99] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, pp. 43–55, 1999.
[Kle02] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acustica united with acta acustica*, accepted (publication scheduled for 2002).
[Mil02] L.M. Miller, M.A. Escabi, H.L. Read, and C.E. Schreiner: "Spectrotemporal Receptive Fields in the Lemniscal Auditory Cortex", *J. Neurophysiol.* 87, pp. 516-527, 2002.
[Sch00] C.E. Schreiner, H.L. Read, and M.L. Sutter: „Modular Organization of Frequency Integration in Primary Auditory Cortex" *Annu .Rev. Neurosc*., 23:501-529, 2000.