Methods for capturing spectro-temporal modulations in automatic speech recognition

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

Summary
Psychoacoustical and neurophysiological results indicate that spectro-temporal modulations play an important role in sound perception. Speech signals, in particular, exhibit distinct spectro-temporal patterns which are well matched by receptive fields of cortical neurons. In order to improve the performance of automatic speech recognition (ASR) systems a number of different approaches are presented, all of which target at capturing spectro-temporal modulations. By deriving secondary features from the output of a perception model the tuning of neurons towards different envelope fluctuations is modeled. The following types of secondary features are introduced: product of two or more windows (sigma-pi cells) of variable size in the spectro-temporal representation, fuzzy-logical combination of windows and a Gabor function to model the shape of receptive fields of cortical neurons. The different approaches are tested on a simple isolated word recognition task and compared to a standard Hidden Markov Model recognition system. The results show that all types of secondary features are suitable for ASR. Gabor secondary features, in particular, yield a robust performance in additive noise, which is comparable and in some conditions superior to the Aurora 2 reference system.

PACS no. 00.00.Xx, 00.00.Xx

1. Introduction

Speech and many other natural sound sources exhibit distinct spectro-temporal amplitude modulations. While the temporal modulations are mainly due to the syllabic structure of speech, resulting in a bandpass characteristic with a peak around 4Hz [1], spectral modulations are due to the harmonic and formant structure of speech. The latter are not at all stationary over time. Coarticulation and intonation result in variations of fundamental and formant frequencies even within a single phoneme (cf. Fig. 1 as an example). The question is whether there is relevant information in amplitude variations oblique to the spectral and temporal axis and how it may be utilized to improve the performance of automatic classifiers.

In automatic speech recognition (ASR) the focus typically is on spectral modulation for a given time frame (cepstral analysis) *and/or* temporal fluctuations in individual frequency channels [2, 3]. Although there are proposals to take two-dimensional variability into account (e.g. [4]), auditory processing is not modeled explicitly.

Therefore, three different approaches are presented in this paper which target at capturing spectro-temporal modulations to increase the robustness of ASR systems:

Sigma-pi cells were originally proposed as a part of ASR systems in order to better capture certain features of speech like formants, formant transitions, fricative onsets and (for larger units) phoneme sequences. A logical ”AND” operation is performed by multiplicative combination of two spectro-temporal windows [5]. A generalization of this approach, towards a larger number of windows and variable window size, is motivated by recent psychoacoustical reverse correlation experiments. Using short segments of semi-periodic white Gaussian noise as stimuli, *early auditory features* of certain spectro-temporal shape were revealed [6]. These findings correspond well to physiological measurements of spectro-temporal receptive fields of neurons in the primary auditory cortex [7] which often encompass different unconnected but highly localized parts of the spectrogram.

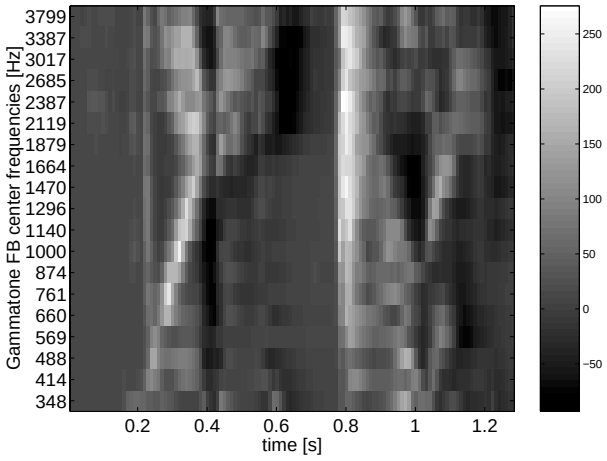

Figure 1. An example of a primary feature matrix for an utterance of the two words ”Woody Allen” - in this case derived from the model of auditory perception as described in Section 3.2. Gray shading denotes output values in model units. A number of diagonal spectro-temporal structures may be identified.

Received 1 January 1995,
accepted 1 January 1995.

Fuzzy logic units: Due to its linear nature, the reverse correlation method does not reveal, if there has to be energy in regions A *and* B in order to stimulate a response or whether the receptive field is simply fragmented. To take account of this ambiguity the sigma-pi cell approach is extended to other fuzzy logical combination of windows, adding OR, NOR and NAND to the multiplicative AND operation.

Gabor functions are localized sinusoids and known to model the receptive fields of certain neurons in the visual system [8]. In addition, experiments on human spectro-temporal modulation perception were modeled well by assuming a response field similar to two-dimensional Gabor functions [9]. Therefore, in the third approach of this paper, two-dimensional Gabor receptive fields are examined for ASR. A complex two-dimensional Gabor function is calculated and reduced to real values by using only the real or imaginary component.

In the following the three types of secondary features are introduced and then applied to a simple isolated word recognition task for a first evaluation. Because of the large number of possible parameter combinations for all three variants of secondary features, the selection of a suitable subset is a major concern and the key to good classification performance. The classification and feature selection scheme described in Sec. 3.3 allows to automatically optimize a subset from all possible secondary features on a given task and is therefore favored over standard ASR back ends in this approach.

2. Secondary features

The secondary features $s_1(t)..s_M(t)$ are calculated from the primary feature values $p(t, f)$, which form a spectro-temporal representation of the input signal. t and f denote time and frequency channel index, respectively. The simplest examples of such two-dimensional representation (amplitude over frequency and time) are the spectrogram obtained by short-term Fourier analysis of consecutive time windows or, alternatively, a bank of bandpass filters. For speech and signal classification purposes, auditory-based approaches are likely to be more appropriate.

2.1. Sigma-pi cells

Sigma-pi cells are known as second order elements from artificial neural network theory. This term describes certain network units in which the weighted outputs from two or more other units are multiplied before summation over all input values.

In the approach presented here, a number of windows $k = 1..K$ are defined centered around one element of the primary feature representation, which is located at frequency channel f_k and by t_k time steps shifted relative to the current feature vector. The windows have the extension Δt_k and Δf_k in time and frequency.

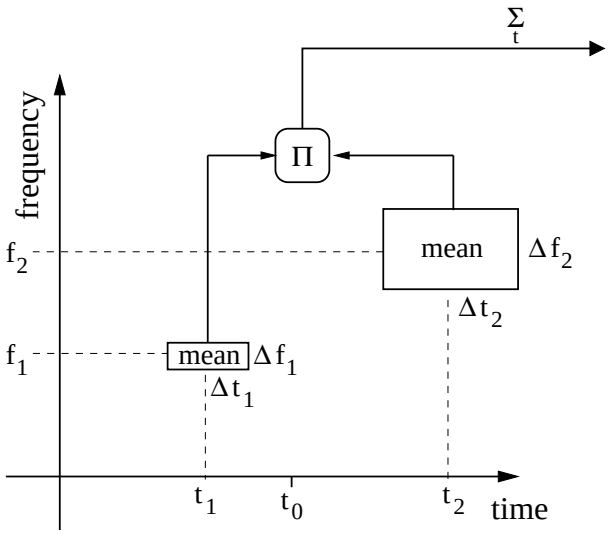

Figure 2. This sketch shows the denotation of parameters for a sigma-pi cell with two windows. See text for further description!

First, the average value w_k of each window is derived by

$$w_k = \frac{1}{\Delta t_k \Delta f_k} \sum_{t'} \sum_{f'} p(t_0 + t_k + t',\ f_k + f') \quad (1)$$

with $-\frac{\Delta t_k}{2} \leq t' \leq \frac{\Delta t_k}{2}$ and $-\frac{\Delta f_k}{2} \leq f' \leq \frac{\Delta f_k}{2}$.

The resulting value of any sigma-pi cell for time frame t_0 is then obtained from the window averages by:

$$s_m(t_k, f_k, \Delta t_k, \Delta f_k, t_0) = \prod_{k=1}^{K} w_k \quad (2)$$

The secondary feature values $s_m(t_0)$ are often averaged over the whole utterance to obtain a single value per sigma-pi cell. Gramß and Strube [5] proposed sigma-pi cells to be used as secondary features based on critical band spectrograms for isolated word recognition. Sigma-pi cells have later been used in combination with a perception model as front end for isolated word recognition and it was shown, that this combination increases the robustness of ASR systems in additive noise [10]. With a non-linear back end the combination of perception model and sigma-pi cells is also suitable for sub-band signal-to-noise ratio (SNR) estimation [11]. In all those applications only two windows were used per sigma-pi cell and the smaller window was restricted to a single element of $p(t, f)$.

In the experiments presented below the window parameters for sigma-pi cells have the following constraints: $t_k = -20..20$ $(-200..200ms)$, $\Delta t_k = 1..10$ $(10..100ms)$, $\Delta f_k = 1..5$ (ERB)[1], and the number of windows $K = 2..3$. Furthermore, the windows have to be non-overlapping. Summation over time is performed to obtain a single secondary feature value per utterance.

[1] equivalent rectangular bandwidth[12]

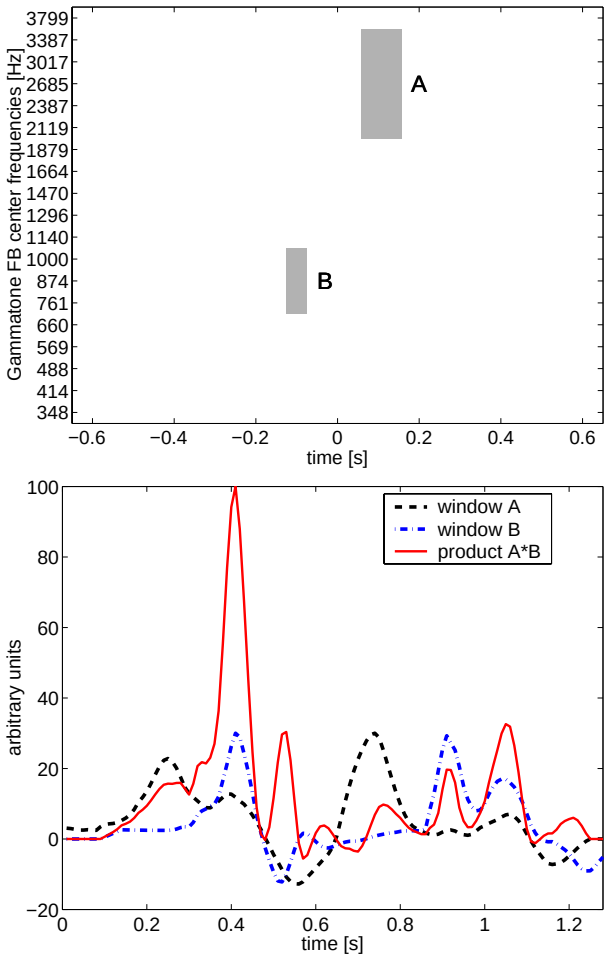

Figure 3.
TOP: An example of a sigma-pi cell with two windows. Window A parameters are: $t = -10$ $(-100ms)$, $f = 7$ (ERB), $\Delta t = 5$ $(50ms)$ and $\Delta f = 3$ (ERB). Window B parameters are: $t = 10$ $(100ms)$, $f = 16$ (ERB), $\Delta t = 10$ $(100ms)$ and $\Delta f = 5$ (ERB).
BOTTOM: Window averages and product of the two windows as a function of time, when the above sigma-pi cells is applied to the utterance depicted in Fig. 1. The combination of the vowels /u/ and /i/ (or the lower and higher formants, respectively) in "Woody" was detected by the sigma-pi cells, by yielding large feature values around 0.4s.

Fig. 3 gives an example on how a sigma-pi may serve as a feature detector. The sigma-pi cell is tuned to a sequence of phonetic elements in that case. The two windows, when coinciding with peaks in the spectro-temporal primary feature representation, basically detect spectro-temporal modulation of the frequency corresponding to the distance between the two windows. The temporal and spectral extension of the windows compensate to some degree for the variability inherent to spoken language. By calculating the product of the two windows, the secondary feature is of second order and the detection information remains even after integration over the whole time span of a word.

2.2. Fuzzy logic units

The sigma-pi cell approach is now extended by using true fuzzy logical combinations of windows instead of a simple multiplication, which corresponds to a logical AND. To obtain a value range between zero and one, the primary feature vectors are normalized by a logistic mapping function over the whole utterance:

$$p'(t,f) = \frac{1}{1+\exp\left[-\frac{p(t,f)-50}{25}\right]}. \tag{3}$$

or, alternatively, by a linear min-max normalization scheme:

$$p'(t,f) = \frac{p(t,f)-\min(p)}{\max(p)-\min(p)}. \tag{4}$$

The window averages w_k are calculated as in Eq. 1. The resulting value of a fuzzy logic unit for time t_0 is obtained recursively by:

$$s_{m,1}(t_0) = W_1(w_1) \tag{5}$$

and

$$s_{m,k}(t_0) = s_{m,k-1}\, O_{k-1}\, W_k(w_k). \tag{6}$$

The recursion terminates after K steps and the value $s_{m,K}$ is than adopted as secondary feature value $s_m = s_{m,K}$ for time t_0. The window operator W_k is either identity ($f(A) = A$) or fuzzy complement (NOT operation), which is defined as $f(A) = 1 - A$. The possible fuzzy operators O_l are

intersection $f(A,B) = min(A,B)$
algebraic product $f(A,B) = A \cdot B$
union $f(A,B) = max(A,B)$
algebraic sum $f(A,B) = A + B - A \cdot B$.

The first two operators represent a fuzzy logical AND while the latter two correspond to fuzzy logical OR. With two or more windows a variety of combinations are possible. The NAND operation ('A AND NOT B'), for example, is assumed to be useful for edge detection in any spectro-temporal direction, while the AND operation ('A AND B', 'A AND NOT B AND C') serves as a detector for spectro-temporal modulations.

In the experiments described below, for fuzzy logic units the same parameter constraints applied as for sigma-pi cells.

2.3. Gabor receptive fields

The receptive field of cortical neurons is modeled as a two-dimensional complex Gabor function $g(t,f)$ defined as the product

$$g(\cdot) = n(\cdot) \cdot e(\cdot) \tag{7}$$

of the Gaussian envelope $n(t,f)$ with parameters $f_0, t_0, \sigma_f, \sigma_t$

$$n(\cdot) = \frac{1}{2\pi\sigma_x\sigma_t} \cdot \exp\left[\frac{-(f-f_0)^2}{2\sigma_f^2} + \frac{-(t-t_0)^2}{2\sigma_t^2}\right] \quad (8)$$

and the complex Euler function $e(t,f)$ with parameters $f_0, t_0, \omega_f, \omega_t$

$$e(\cdot) = \exp\left[i\omega_f(f-f_0) + i\omega_t(t-t_0)\right] \quad (9)$$

by using either the real or imaginary component only. The envelope width is defined by standard deviation values σ_f and σ_t. These are chosen as $\sigma = \frac{1}{\omega} \implies \sigma = \frac{T}{2\pi}$ for the imaginary component to ensure that only one period of the oscillation gives a significant contribution to the function, and as $\sigma = \frac{\pi}{\omega} \implies \sigma = \frac{T}{2}$ for the real component. In the latter case the chosen combination of spread and periodicity leads to about 2.5 periods of the oscillation in the envelope and results in a negligible bias because

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\cdot)\,\mathrm{dt}\,\mathrm{df} \le \exp\left[-\frac{\omega_t^2\sigma_t^2 + \omega_x^2\sigma_x^2}{2}\right] \quad (10)$$

and, with $\sigma_t = \frac{\pi}{\omega_t}$ and $\sigma_f = \frac{\pi}{\omega_f}$,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\cdot)\,\mathrm{dt}\,\mathrm{df} \le \exp\left[-\pi^2\right]. \quad (11)$$

This is important, because otherwise any stationary background signal would contribute to the secondary feature value.

In the experiments below the allowed temporal modulation frequencies $\frac{\omega_t}{2\pi}$ are limited to a range of one to 30Hz and the spectral modulations $\frac{\omega_f}{2\pi}$ to a range of 0.05 to 0.3 cycl/ERB, roughly corresponding to 0.25 - 1.5 cycl/oct. For a one ERB spectral resolution of the primary features, spectral modulations may only be calculated up to 0.5 cycles/ERB.

In order to extract a secondary feature value, the correlation between Gabor receptive field and the primary feature matrix is calculated. This matched filter operation is carried out in each frequency channel and the resulting values are summarized over all channels to obtain the activation $a(t_0, f_0, \omega_f, \omega_t, \sigma_f, \sigma_t)$ for each time step t_0. The cell response or secondary feature value for the whole utterance is then calculated as follows:

$$s_m(f_0, \omega_f, \omega_t, \sigma_f, \sigma_t) = \sum_{t_0=1}^{T} T\left[\, a(t_0)\,\right] \quad (12)$$

with the non-linear transformation function T by either full-wave or half-wave rectification of $a(t_0)$.

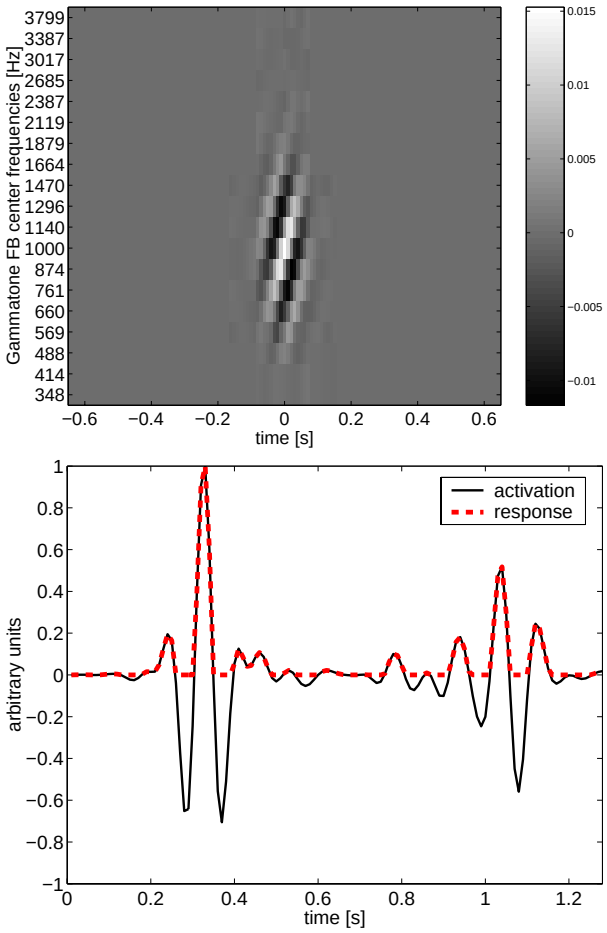

Figure 4.
TOP: Example of the real component of a 2D Gabor function spectrally centered at 1000 Hz. Function values are given in shadings of gray. The Euler frequencies are $\frac{\omega_t}{2\pi} = -12Hz$ and $\frac{\omega_f}{2\pi} = 0.2$cycles/channel. The function is calculated on a grid with 100 Hz temporal and 1/ERB spectral sampling, according to the primary feature extraction method used in this study.
BOTTOM: Filter output ("activation") and halfway rectified feature values ("response") over time when the above Gabor filter is applied to the utterance depicted in Fig. 1. The rising formant between 0.3 and 0.4s fits the Gabor filter shape well and yields highest feature values. A similar diagonal feature is detected around 1.1s, resulting in a second, somewhat smaller peak.

In the experiments presented below, the primary feature vector sequence $p(t,f)$ is used either without or with min-max normalization (Eq. 4).

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. The wide range of possible Gabor features is therefore versatile enough to contain purely spectral features (as cepstra) or temporal processing (as in the RASTA or TRAPS approaches). The above mentioned front ends are extended as most of the possible Gabor filters perform integrated spectral *and* temporal processing. Fig. 4 shows one example of such a diagonal

Gabor feature function and how it can be used to detect formant transitions.

3. Automatic speech recognition experiments

3.1. Material

The speech material for training and testing is taken from the ZIFKOM database[2]. Each German digit was recorded once from 200 different speakers. The speech material is equally divided into two parts for training and testing, each consisting of 1000 utterances by 50 male and 50 female speakers. Training is performed on clean digits only. Testing is performed on clean and on noisy digits. For distortion, three types of noise are added to the utterances with SNR between 25 and -5dB: a) un-modulated speech shaped noise (CCITT G.227), with a spectrum similar to the long-term spectrum of speech, b) real babble noise recorded in a cafeteria situation and c) speech-like shaped and modulated noise (ICRA noise signal 7, [13])[3]. Before mixing, speech and noise signals are bandpass filtered to 300-4000Hz, roughly corresponding to the telephone band.

3.2. Primary feature extraction

The output of the model of auditory perception (PEMO) is used as primary feature matrix. PEMO has been originally developed by Dau et al. [14] for quantitatively simulating psychoacoustical experiments, such as temporal and spectral masking, and has been successfully applied as a robust front end in isolated word recognition experiments [15, 16]. Its major components are the peripheral gammatone filter bank [17] and the non-linear adaptation loops [18], which perform a log-like compression for stationary signals and emphasize onsets and offsets of the envelope. This causes a sparse coding of the input in the spectro-temporal domain. It should be stressed, that any other time-frequency amplitude representation could also be used with this approach, preferably an auditory model or auditory-like processing [11].

In this study, the model was slightly modified by adding a pre-emphasis[4], which is motivated by earlier ASR experiments [10]. Overall, 19 frequency channels are used with bandwidth and spacing of one ERB and center frequencies ranging from 384 to 3799Hz. The primary feature vectors are then derived by downsampling the model output to a sampling frequency of $f_s = 100$Hz in each channel.

3.3. Recognizer

For classification and optimization of the type of secondary features the *Feature-finding Neural Network* (FFNN) [5] is used. It consists of a linear single-layer perceptron in conjunction with secondary feature extraction and an optimization rule for the feature set. For a sufficiently high-dimensional feature space (i.e. a large number of secondary features), a linear net should classify equally well as non-linear classifiers and fast training is guaranteed by matrix inversion (pseudo-inverse method). Given P examples, each represented by a secondary feature vector with M elements, the feature vectors form a $M \times P$ feature matrix $\mathbf{X}$. Given the target matrix $\mathbf{Y}$ ($N \times P$ with N as the number of classes or target values per example), the optimal (in RMS sense) weight matrix $\mathbf{W}$ ($N \times M$) is found analytically by calculating the pseudo-inverse

$$\mathbf{X}^+ = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \tag{13}$$

of the secondary feature matrix $\mathbf{X}$. The weight matrix is obtained as

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^+ \tag{14}$$

and minimizes the classification error

$$E = |\mathbf{Y} - \mathbf{W}\mathbf{X}|^2. \tag{15}$$

Gramß [19] proposed a number of training algorithms for the FFNN system, one of which, the *substitution rule*, is used in this study:

1. Choose M secondary features arbitrarily.
2. Find the optimal weight matrix $\mathbf{W}$ using all M features and the M weight matrices that are obtained by using only $M-1$ features, thereby leaving out every feature once.
3. Measure the relevance R of each feature i by
$$R_i = E(\text{ without feature } i) - E(\text{ with all features}) \tag{16}$$
4. Discard the least relevant feature $j = \text{argmin}(R_i)$ from the subset and randomly select a new candidate.
5. Repeat from point 2. until the maximum number of iterations is reached.
6. Recall the set of secondary features, that performed best on the training / validation set and return it as result of the substitution process (modification from original substitution rule).

Although the classification is performed by a linear neural network, the whole classification process is highly non-linear due to the second order characteristics of the secondary features. The thereby obtained set of secondary features might also be used as input to other, more sophisticated classification systems. The segmentation problem is not relevant for an isolated word recognition task and therefore the summation of secondary feature values over the whole utterance is a sufficiently good option to derive a single value per secondary feature and utterance. In the more general continuous case, e.g., a leaky integrator could be used to extract time-depending secondary feature values.

[2] Deutsche Telekom AG

[3] two foreground speakers and four background speakers

[4] differentiation with factor of 0.97: $y_n = x_n - 0.97 \cdot y_{n-1}$

In the experiments below, a set of 60 secondary features is optimized over 2000 iterations. Due to the non-deterministic nature of the substitution rule (random start set and randomly chosen substituting secondary feature), training is carried out eight times per configuration.

3.4. Results

The results are summarized in Tab. I. All three types of secondary feature are suitable for ASR. Gabor features perform best in CCITT noise and on clean test material and comparable to sigma-pi cells for babble and ICRA 7 noise. Fuzzy logic secondary features lead to an unacceptable high error for clean test data and also to the highest word error rate (WER) values in most other cases. The robustness of fuzzy logic features can be increased by using min-max normalization instead of logistic function (Tab. II), but the error rate for clean data remains too high also in that case.

Table I. Word error rates (WER) in percent for different SNR (in dB) and noise conditions. 'train' indicates the training material, while 'clean' refers to the unmixed test data. Mean and standard deviation (in brackets) over 8 training runs per condition are given for sigma-pi cell, fuzzy logic (logistic normalization) and gabor secondary features.

cond.	SNR	Sigma-pi		Fuzzy (logistic)		Gabor	
train		0.5	(0.2)	1.0	(0.2)	0.4	(0.2)
clean		2.0	(0.3)	3.3	(0.6)	1.1	(0.2)
ccitt	25	4.9	(1.2)	9.0	(2.7)	5.1	(1.2)
	20	11.7	(2.0)	22.2	(7.4)	11.1	(3.1)
	15	35.3	(4.0)	47.9	(10.9)	27.5	(8.6)
	10	67.1	(4.8)	72.3	(6.1)	52.7	(9.9)
	5	82.8	(5.2)	83.5	(3.4)	72.0	(5.5)
	0	88.5	(1.7)	88.2	(1.2)	82.3	(3.8)
	-5	89.8	(0.3)	89.6	(0.5)	87.2	(2.3)
babble	25	3.6	(0.7)	8.2	(0.8)	4.5	(1.0)
	20	6.3	(1.6)	16.3	(2.3)	8.6	(2.7)
	15	16.9	(3.5)	33.5	(7.8)	22.2	(7.4)
	10	43.0	(4.7)	54.5	(11.2)	45.8	(10.5)
	5	68.1	(4.6)	72.0	(7.8)	68.0	(9.0)
	0	82.4	(3.9)	82.1	(3.4)	81.3	(4.8)
	-5	87.4	(2.4)	87.5	(2.3)	87.5	(2.1)
icra7	25	3.6	(0.7)	7.4	(1.1)	4.0	(1.1)
	20	6.6	(1.3)	15.1	(3.4)	9.0	(4.0)
	15	17.2	(4.1)	30.7	(7.1)	23.5	(11.7)
	10	44.5	(6.3)	51.6	(9.8)	46.1	(18.3)
	5	70.9	(3.2)	70.5	(8.0)	66.4	(17.5)
	0	83.0	(2.5)	80.9	(5.3)	78.3	(12.8)
	-5	87.9	(1.9)	86.4	(2.7)	84.3	(7.4)

Gabor receptive fields yield lower WER values than sigma-pi cells in most cases. This is remarkable, because the Gabor secondary features are of 1st order, while the other two variants are 2nd order features. The variance of performance over different training runs is relatively high, especially for Gabor receptive fields in the case of additive speech-like modulated noise (ICRA 7). As the optimization is carried out on clean training data, only in some cases the secondary features seem to be affected by the modulation in the noise signal (which is kept frozen for all examples). In Tab. II WER for the most robust single set of Gabor features out of eight sets are shown. The large variance of WER in noise between the eight sets of optimized Gabor secondary features indicate, that that some sets of Gabor receptive fields contain features which are less suitable in noisy conditions. Multi-condition training is likely to increase the robustness by selecting only noise-robust type of features into the optimal set.

Table II. Word error rates (WER) in percent for different SNR (in dB) and noise conditions. 'train' indicates the training material, while 'clean' refers to the unmixed test data. Mean and standard deviation (in brackets) over 8 training runs per condition are given for fuzzy logic units and Gabor receptive fields - both with min-max normalization of primary feature vectors. The most robust single set of gabor features without normalization ('Gab. best') is compared to the Aurora 2 baseline system ('Aurora'), which is given as a reference.

cond.	SNR	Fuzzy (min-max)		Gabor (min-max)		Gab. best	Aurora
train		0.5	(0.1)	0.3	(0.1)	0.5	0.3
clean		3.7	(0.6)	1.7	(0.3)	1.1	0.3
ccitt	25	4.6	(1.0)	3.8	(0.8)	4.6	1.7
	20	6.8	(1.4)	5.8	(1.8)	7.6	3.9
	15	12.9	(2.9)	12.0	(4.1)	16.7	9.7
	10	29.2	(6.2)	26.8	(8.5)	37.9	24.1
	5	51.8	(9.2)	50.1	(11.5)	66.4	73.8
	0	69.8	(8.2)	73.0	(10.0)	80.5	90.9
	-5	81.3	(5.7)	85.4	(5.3)	85.0	90.6
babble	25	4.4	(0.6)	3.4	(0.5)	3.4	1.2
	20	6.1	(1.0)	5.2	(1.1)	4.8	2.3
	15	10.7	(1.1)	10.3	(2.5)	9.0	4.1
	10	21.9	(2.5)	22.4	(5.2)	22.7	14.1
	5	42.5	(5.1)	43.4	(5.5)	46.6	42.0
	0	65.9	(6.6)	64.9	(6.0)	70.3	72.6
	-5	82.0	(4.9)	80.0	(3.5)	83.0	83.5
icra7	25	4.6	(0.9)	2.8	(0.5)	2.8	1.1
	20	7.6	(1.4)	4.7	(0.9)	3.8	1.6
	15	14.6	(2.2)	9.4	(2.8)	7.4	4.0
	10	28.3	(3.9)	20.3	(6.0)	15.5	14.8
	5	48.0	(4.6)	38.9	(7.6)	30.2	31.3
	0	67.8	(2.6)	59.8	(5.3)	50.7	54.8
	-5	81.3	(2.3)	75.6	(4.9)	69.1	83.7

As a reference, the Aurora 2 baseline system [20] has been applied to the same classification task. It is composed out of the WI007 (mel-cepstrum) front end and a reference HTK recognizer. The results obtained by this Hidden Markov Model classifier are presented in Tab. II and compared to improved Gabor secondary features.

Both, the best Gabor set of secondary features and Gabor secondary feature set with min-max normalization of primary feature values, show a comparable robustness to the aurora baseline system on the given classification task. There is a trend for the aurora system to yield lower WER for clean test data and high SNR values of over 10dB while the Gabor secondary features seem to be superior in more unfavorable conditions of low SNR values. It should be stressed, that the classifier used here for the secondary features is as simple as possible with a summation over the whole utterance followed by a linear neural network. Therefore, an increase in performance can be expected when combining time-dependent secondary features, e.g., Gabor receptive fields, with a more sophisticated classifier.

4. Discussion

The proposed extensions to the secondary feature approach are all suitable for for robust isolated word recognition. Especially the Gabor receptive field method seems to be worthwhile to be investigated further. Gabor secondary features combined with a simple linear classifier show a comparable performance to the state-of-the-art Aurora 2 HMM system. They can be assumed to have a large potential. Earlier studies indicate, for example, an increase in robustness equivalent to a five to eight dB effective gain in SNR by using noise reduction pre-processing schemes with PEMO primary features [16]. Classification performance should increase further by replacing the simple linear network classifier with a state-of-the-art HMM back end and/or adding spectro-temporal features as another feature stream in a multi-stream system.

Acknowledgement

The author would like to thank Volker Hohmann and Birger Kollmeier for their substantial support and contribution to this work. Thanks also to Christian Kaernbach for stimulating conversation and his idea to use fuzzy logic, to Heiko Gölzer for fruitful discussion about optimization rules.
This work was supported by *Deutsche Forschungsgemeinschaft* (Project ROSE, Ko 942/15-1).

References

[1] N. Kanedera, T. Arai, H. Hermansky, M. Pavel: On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication **28** (1999) 43–55.
[2] H. Hermansky, N. Morgan: RASTA processing of speech. IEEE Trans. Speech Audio Processing **2** (1994) 578–589.
[3] H. Hermansky, S. Sharma: TRAPS - Classifiers of temporal patterns. Proc. ICSLP'98, 1998. 1003–1006.
[4] K. Weber, S. Bengio, H. Bourlard: HMM2 - A novel approach to HMM emission probability estimation. ICSLP, 2000.
[5] T. Gramß, H. W. Strube: Recognition of isolated words based on psychoacoustics and neurobiology. Speech Communication **9** (1990) 35–40.
[6] C. Kaernbach: Early auditory feature coding. Contributions to psychological acoustics: Results of the 8th Oldenburg Symposium on Psychological Acoustics., 2000. BIS, Universität Oldenburg, 295–307.
[7] R. C. deCharms, D. T. Blake, M. M. Merzenich: Optimizing sound features for cortical neurons. Science **280** (1998) 1439–1443.
[8] R. De-Valois, K. De-Valois: Spatial vison. Oxford U.P., New York, 1990.
[9] T. Chi, Y. Gao, M. C. Guyton, P. Ru, S. Shamma: Spectro-temporal modulation transfer functions and speech intelligibility. J. Acoust. Soc. Am. **106** (1999) 2719–2732.
[10] M. Kleinschmidt, V. Hohmann: Perzeptive Vorverarbeitung und automatische Selektion sekundärer Merkmale zur robusten Spracherkennung. Fortschritte der Akustik, DAGA Oldenburg, 2000. DEGA, 382–383.
[11] M. Kleinschmidt, V. Hohmann: Sub-band SNR estimation using auditory feature processing. Speech Communication (2002). Special Issue on Digital Hearing Aids (submitted).
[12] B. C. J. Moore, B. R. Glasberg: Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. **74** (1983) 750–753.
[13] International Collegium of Rehabilitory Audiology (ICRA) - Hearing Aid Clinical Test Environment Standardization Work Group: ICRA noise signals, version 0.3. CDROM, 1997.
[14] T. Dau, D. Püschel, A. Kohlrausch: A quantitative model of the 'effective' signal processing in the auditory system: I. Model structure. J. Acoust. Soc. Am. **99** (1996) 3615–3622.
[15] J. Tchorz, B. Kollmeier: A model of auditory perception as front end for automatic speech recognition. J. Acoust. Soc. Am. **106** (1999) 2040–2050.
[16] M. Kleinschmidt, J. Tchorz, B. Kollmeier: Combining speech enhancement and auditory feature extraction for robust speech recognition. Speech Communication **34** (2001) 75–91. Special Issue on Robust ASR.
[17] V. Hohmann: Gammatone filter bank and re-synthesis. Acustica united with Acta Acustica (2001). This issue.
[18] D. Püschel: Prinzipien der zeitlichen Analyse beim Hören. Doctoral thesis, Universität Göttingen, 1988.
[19] T. Gramß: Fast algorithms to find invariant features for a word recognizing neural net. IEEE 2nd International Conference on Artificial Neural Networks, Bournemouth, 1991. 180–184.
[20] H. Hirsch, 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.