ROBUST SPEECH RECOGNITION BASED ON LOCALIZED SPECTRO-TEMPORAL FEATURES

Bernd Meyer and Michael Kleinschmidt

Medizinische Physik, Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg
{Bernd.Meyer; Michael.Kleinschmidt}@uni-oldenburg.de
http://medi.uni-oldenburg.de/projects/asr

Abstract: In order to enhance automatic speech recognition performance in adverse conditions, localized spectro-temporal features (LSTF) are investigated, which are motivated by physiological measurements in the primary auditory cortex. In the Aurora2 experimental setup, Gabor-shaped LSTFs combined with a Tandem system yield robust performance with a feature set size of 30. If computational constraints allow, the set size may be increased with some beneficial effect to up to 70 features. There is supportive evidence that the previously chosen 1.5 periods within the envelope yield are a reasonable choice. Improved results can be obtained when using a Hanning window instead of a cut-off Gaussian envelope due to better modulation frequency characteristics. Combined spectro-temporal modulations filters play an important role in characterizing speech as more than 40% of all automatically selected features exhibit diagonal characteristics.

1 INTRODUCTION

The large gap in performance between normal-hearing native listeners and state-of-the art ASR systems is most drastically encountered in adverse acoustic conditions and prohibits automatic speech recognition (ASR) technology from being widely used. Consistently, humans outperform machines by at least an order of magnitude [1]. Human listeners recognize speech even in very adverse acoustical environments with strong reverberation and interfering sound sources. While many cognitive aspects of speech perception still lie in the dark, there is much progress in the research on signal processing in the more peripheral parts of the (human) auditory system. Our work is thus led by the idea of learning certain feature extraction techniques from the biological blueprint.

Recent findings from a number of physiological experiments in different mammal species showed that a large percentage of neurons in the primary auditory cortex (A1) respond differently to upward- versus downward-moving ripples in the spectrogram of the input [2]. Individual neurons are sensitive to specific spectro-temporal modulation frequencies in the incoming sound signal. The spectro-temporal receptive fields (STRF) often clearly exceed one critical band in frequency, have multiple peaks and also show tuning to temporal modulation. Still, the STRF patterns are mainly localized in time and frequency, generally spanning at most 250 ms and one or two octaves, respectively.

The neurophysiological data fit well with psychoacoustic experiments on early auditory features: in [3] a psychophysical reverse correlation technique was applied to masking experiments with semi-periodic white noise. The resulting basic auditory feature patterns are distributed in time and frequency and in some cases comprised of several unconnected parts, very much resembling the STRF of cortical neurons.

In the visual cortex, STRFs are measured with (moving) orientated grating stimuli. The results very well match two-dimensional Gabor functions [4]. The use of 2D complex Gabor filters as features for ASR has been proposed earlier and proven to be relatively robust as part of a high end system [5]. This novel approach of spectro-temporal processing by using localized sinusoids most closely matches the neurobiological data and also incorporates other features as special cases: purely spectral Gabor functions perform sub-band cepstral analysis—modulo the windowing function—and purely temporal ones can resemble TRAPS or the RASTA impulse response and its derivatives [6] in terms of temporal extent and filter shape.

In this paper, the localized spectro-temporal features (LSTF) are analyzed further with respect to their envelope shape and parameter constrains. This includes an analysis of the variability inherent to the feature selection procedure and investigation regarding the number of features needed for robust ASR system performance. In addition, alternative envelope characteristics of the filter impulse response are compared and evaluated.

2 Localized spectro-temporal features

A spectro-temporal representation of the input signal is processed by a number of 2-D modulation filters. The filtering is performed by correlation over time of each input frequency channel with the corresponding part of the LSTF function (centered on the current frame and desired frequency channel) and a subsequent summation over frequency. This yields one output value per frame per filter and is equivalent to a 2-D correlation of the input representation with the complete filter function and a subsequent selection of the desired frequency channel of the output. In this study, log mel-spectrograms serve as input features for feature extraction. This was chosen for its widespread use in ASR and because the logarithmic compression and mel-frequency scale might be considered a very simple model of peripheral auditory processing. Any other spectro-temporal representation of speech could be used instead and especially more sophisticated auditory models might be a good choice for future experiments.

The two-dimensional complex Gabor function $g(n, k)$ as proposed in [7] for ASR is defined as the product of a Gaussian envelope $g(n, k)$ and the complex sinusoidal function $s(n, k)$ (c.f. Fig. 1 a and c). The envelope width is defined by standard deviation values σ_n and σ_k, while the periodicity is defined by the radian frequencies ω_n and ω_k with n and k denoting the time and frequency index, respectively. The two independent parameters ω_n and ω_k allow the Gabor function to be tuned to particular directions of spectro-temporal modulation, including *diagonal* modulations. Further parameters are the centers of mass of the envelope in time and frequency n_0 and k_0. In this notation the Gaussian envelope $g(n, k)$ is defined as

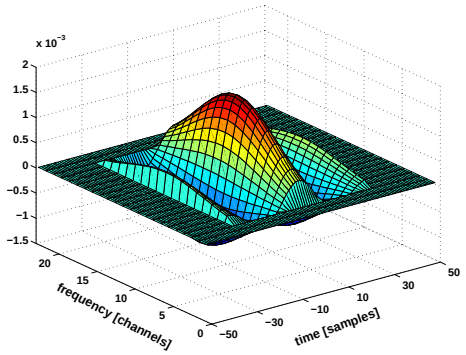
$$g(n,k) = \frac{1}{2\pi\sigma_n\sigma_k} \cdot \exp\left[\frac{-(n-n_0)^2}{2\sigma_n^2} + \frac{-(k-k_0)^2}{2\sigma_k^2}\right] \tag{1}$$

and the complex sinusoid $s(n, k)$ as

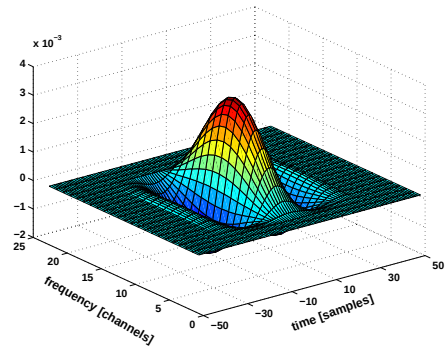
$$s(n,k) = \exp\left[i\omega_n(n-n_0) + i\omega_k(k-k_0)\right]. \tag{2}$$

The envelope width is chosen depending on the modulation frequency ω_x, respective the corresponding period T_x, either with a fixed ratio $\nu_x = T_x/2\sigma_x$ = 1 to obtain a 2D wavelet prototype or by allowing a certain range $\nu_x = 1..3$ with individual values for T_x being optimized in the automatic feature selection process. The infinite support of the Gaussian envelope is cut off at $1.5\sigma_x$ from the center. For time dependent features, n_0 is set to the current frame, leaving k_0, ω_k and ω_n as free parameters. From the complex results of the filter operation, real-valued features may be obtained by using the real or imaginary part only. In this case, the overall DC bias was

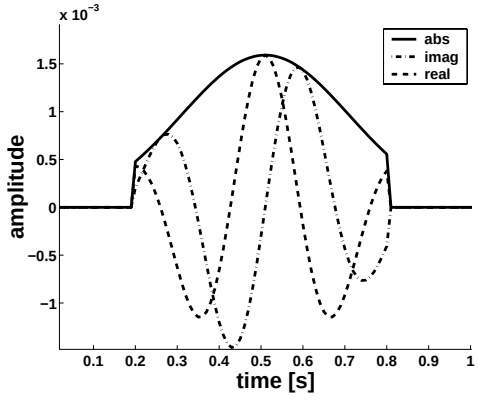
a) Real part of Gabor LSTF (Gaussian envelope)



b) Real part of LSTF (Hanning envelope)



c) Complex Gabor LSTF (Gaussian envelope)



d) Complex LSTF (Hanning envelope)

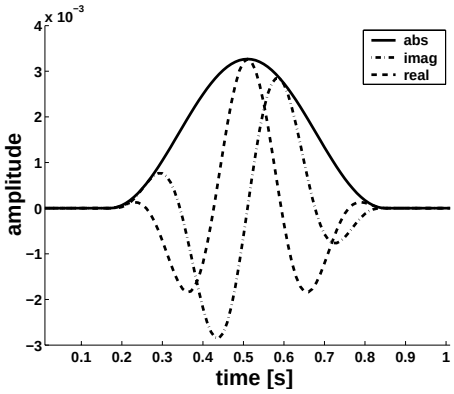


Figure 1 - Illustration of 1- and 2-dimensional filter prototypes for LSTFs with Gabor envelope (left panel, support reduced to $[-1.5\sigma \;\; 1.5\sigma]$) and Hanning envelope (right panel). In the top row the real part of complex 2D impulse responses is depicted. The bottom row show real and imaginary parts as well as envelope of one dimensional LSTFs, corresponding to a cross section of a two dimensional LSTF.

removed from the template. The magnitude of the complex output can also be used. Special cases are temporal filters ($\omega_k = 0$) and spectral filters ($\omega_n = 0$). In these cases, σ_x replaces $\omega_x = 0$ as a free parameter, denoting the extent of the filter, perpendicular to its direction of modulation.

Alternatively, the filter can be designed as the product of a Hanning envelope $h(n,k)$

$$h(n,k) = 0.5 \; - \; 0.5 \cdot \cos\left(\frac{2\pi(n-n_0)}{W_n+1}\right) \cdot \cos\left(\frac{2\pi(k-k_0)}{W_k+1}\right). \tag{3}$$

and the sinusoidal function $s(n,k)$ as above, yielding the window lengths W_n and W_k as parameters instead of σ_n and σ_k (c.f. Fig. 1 b and d).

3 Feature set optimization

The main problem of LSTF is the large number of possible parameter combinations. This issue may be solved implicitly by automatic learning in neural networks with a spectrogram input and a long time window of e.g. 1 s. However, this is computationally expensive and prone to overfitting, as it requires large amounts of training data, which are often unavailable. By putting further constraints on the spectro-temporal patterns, the number of free parameters can be decreased by several orders of magnitude. This is the case when a specific analytical function,

such as the Gabor function [7], is explicitly demanded. Neurophysiological and psychoacoustic knowledge can be exploited for the choice of the prototype, as it is done here.

Feature set optimization is carried out by a modified version of the Feature-finding Neural Network (FFNN). It consists of a linear single-layer perceptron in conjunction with an optimization rule for the feature set [8]. 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 [9] requires iterative re-training of the classifier and replacing the least relevant feature in the set with a randomly drawn new one. When the linear network is used for digit classification without frame by frame target labeling, temporal integration of features is carried out 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 digit recognition in combination with Gabor features in the past [7, 10].

4 Experiments and Results

4.1 Experimental setup

From American English digits strings (TIDigits corpus) and a set of LSTF prototypes, secondary features were computed according to Section 2 and fed into a tandem recognition system [11]. The 60-dimensional feature vector is online normalized and combined with delta and double-delta derivatives before feeding into the MLP[1] (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 a fixed HTK back, configured according to the Aurora 2 experimental framework. The HTK was trained on multicondition or clean only training data (see [12] for details).

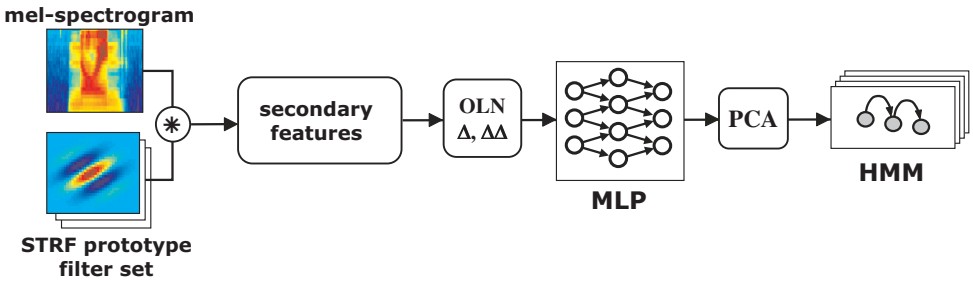


Figure 2 - Schematic overview of the experimental setup. Feature vectors are obtained from correlation of mel-spectrograms with LSTF prototypes and fed into a Tandem recognition system. See text for further description.

In the first two experiments (4.2 and 4.3) features were computed using the set G3 from [5] which was optimized on noisy German digits (ZIFKOM corpus). G3 yields improvements of over 50 % compared to the baseline for clean training in a single stream experiment and improvements of 36 % and 74 % for noisy and clean training, respectively, in a multi-stream combination with the Qualcomm-ICSI-OGI front end [13]

[1] QuickNet software package provided by ICSI, http://www.icsi.berkeley.edu

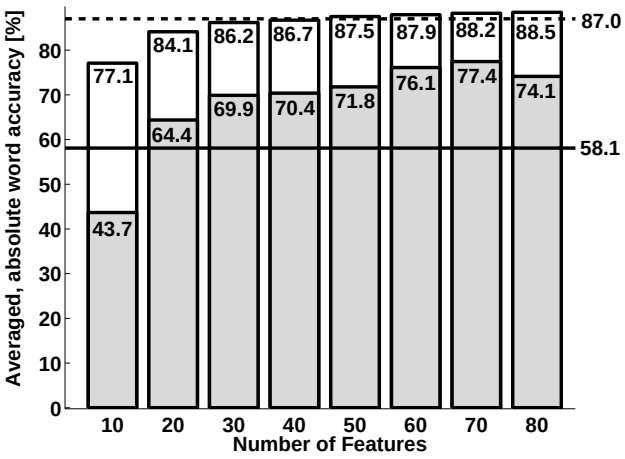


Figure 3 - Averaged recognition performance for different number of features: results are shown for clean condition training (grey) and multi condition training (white). Baseline results are plotted as horizontal lines for multi condition training (dashed) and clean condition training (solid).

4.2 Optimal number of features

Higher number of features require more computation time and do not necessarily lead to improved recognition performance. It is therefore desirable to determine the optimal number of LSTFs. In this experiment the number of features used as input for the tandem system was varied from 10 to 80 features. A reduction of number of features would result in fewer input neurons for the MLP, thus decreasing the total number of weights. For a fair comparison of classification performance, the number of neurons in the hidden layer was adjusted, so that the total number of weights remained constant at about 180,000. The feature set G3, which was used in this experiments, consists of 80 feature prototypes ordered by relevance. When using less than 80 features, the most relevant prototypes were chosen.

In Fig. 3 the obtained recognition rates are shown. While recognition rates for multi condition training steadily increase with higher number of features, this is not the case for clean condition training, where performance drops when using 80 instead of 70 features. However, both curves show saturation at 60 features, while performance superior to the baseline results is already achieved with 50 features for multi-condition training and 20 features for clean-condition training. The optimal number of features in the set would depend on application restrictions. Acceptable performance is reached with as few as 30 and optimal performance with 70 features for set G3. The decrease of word accuracy from 70 to 80 features indicates that the least important 10 features in the set even have a detrimental affect on recognition performance, possibly a result of the optimization algorithm (c.f. Section 3)

4.3 Comparison of envelope widths

The LSTF prototypes in set G3 show more than one maximum because the interval $[-\sigma_x \;\; \sigma_x]$ was chosen to contain exactly one period ($\nu_x = 1$). Still, the support was cut off at 1.5σ, leading to secondary maxima. However, in neurophysiological STRFs commonly only one maximum ist observed. In order to investigate the influence of envelope width, a new feature sets was produced by modifying the existing feature set G3: Halving the values for σ_n and σ_k yields feature set G3sn, where the number of maxima within the Gaussian envelope is limited to one. Using this set, secondary features were computed and fed into the Tandem system.

Table 4.3 shows that the new set G3sn performs worse than G3 for clean training condition,

	average absolute accuracy [%]				relative improvement [%]		
training condition	multi		clean		multi		clean
test condition	all	clean	all		all	clean	all
baseline	87.03	98.66	58.06		0.00	0.00	0.00
G3	87.92	98.17	76.11		2.90	-38.01	50.05
G3sn	87.60	98.82	71.39		3.69	11.46	36.91

Table 1 - Absolute word recognition rates and error reduction relative to baseline for set G3sn with halved envelope width compared to baseline results and feature set G3.

while it yields improved results for multi condition training and clean or high SNR test conditions. Thus, it performs better for matched and worse for missmatched conditions.

4.4 Envelope optimization

Cutting off the support of the Gaussian envelope at $1.5\,\sigma$ as shown in figure 1 results in unwanted higher harmonic frequencies in the modulation frequency domain. These distortions can be eliminated to a great extent by replacing the Gaussian envelope with a Hanning window. Fig. 4 shows a comparison of the spectro-temporal modulation transfer function of the two filter types.

a) LSTF with cut-off Gaussian envelope

b) LSTF with Hanning envelope

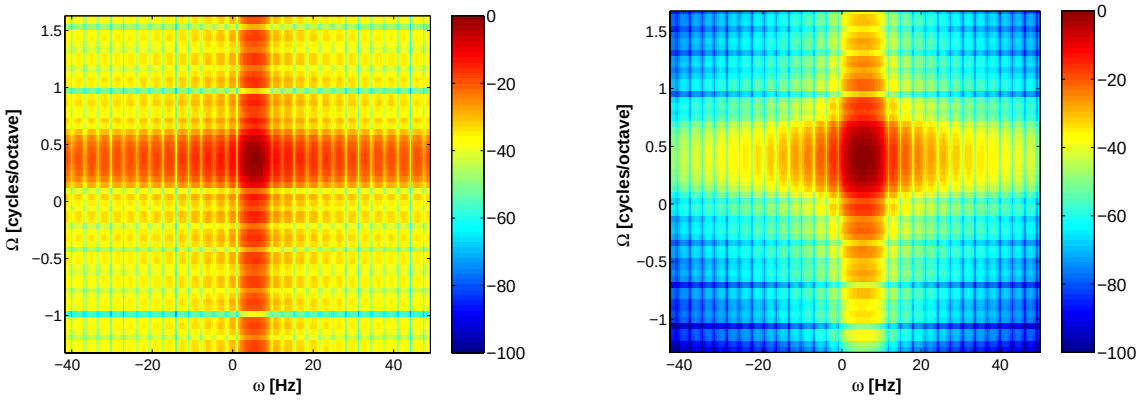


Figure 4 - Absolute values of spectro-temporal transfer functions for real part of LSTF prototypes plotted on logarithmic scale. The shading denotes the amplitude in dB.

In order to determine if the favorable modulation frequency characteristics of Hanning envelopes lead to improved recognition performance, eight feature sets with Gaussian and eight feature sets with Hanning envelope were generated by the automatic optimization procedure (Section 3) with ZIFKOM German digit data. Temporal and spectral modulation frequencies were randomly chosen in an interval from 2 to 50 Hz and 0.06 to 0.5 cycles/octave, respectively. The width of the envelope was loosely coupled to the modulation frequency ω_x, using a value from 1 to 3 for the number of periods ν_x that lie in the interval $[-\sigma_x \;\; \sigma_x]$ for Gaussian envelopes or in the interval $[-W_x/1.5 \;\; W_x/1.5]$ for Hanning envelopes. Boundary conditions for ν_x guaranteed that even at low modulation frequencies the extension of the prototypes did not exceed 23 frequency channels or 101 time frames (corresponding to 1 second filter length). Either absolute, imaginary or real part of the filter output were used as features. German digits (ZIFKOM) mixed with different noise conditions were used for optimization. Each set contained 80 feature prototypes, from which the most relevant 60 were used in the experiment.

	average absolute accuracy [%]		relative improvement [%]	
	multi	clean	multi	clean
a) baseline	87.0	58.1	0.00	0.00
b) G3	87.9	76.1	2,9	50.05
c) Avg Hanning	87.71 ± 0.24	78.43 ± 1.4	$1{,}14 \pm 4.55$	53.53 ± 3.13
d) Avg Gauss	86.78 ± 0.40	76.33 ± 2.11	$-3{,}09 \pm 4.76$	$49{,}97 \pm 4.66$
e) Hanning HB02	88.00	80.51	7,93	58,83
f) Gauss GB07	87.4	76.1	2,55	49.62
g) Gauss GB03	86.9	80.4	-0,15	56.7

Table 2 - Recognition accuracy and relative reduction of error compared to the baseline for different feature types. Beside the baseline data (a), results are shown for feature set G3 (b), averaged values with standard deviation for eight Hanning and eight Gaussian envelope sets (c & d) and best Hanning and Gaussian envelope sets (e) - (g)

The results in Tab. 2 show that in average Hanning-shaped LSTFs outperform Gabor-shaped features in all conditions. The best feature set with Hanning envelope HB02 also outperforms the reference feature set G3 and the best LSTF set with Gaussian envelope.

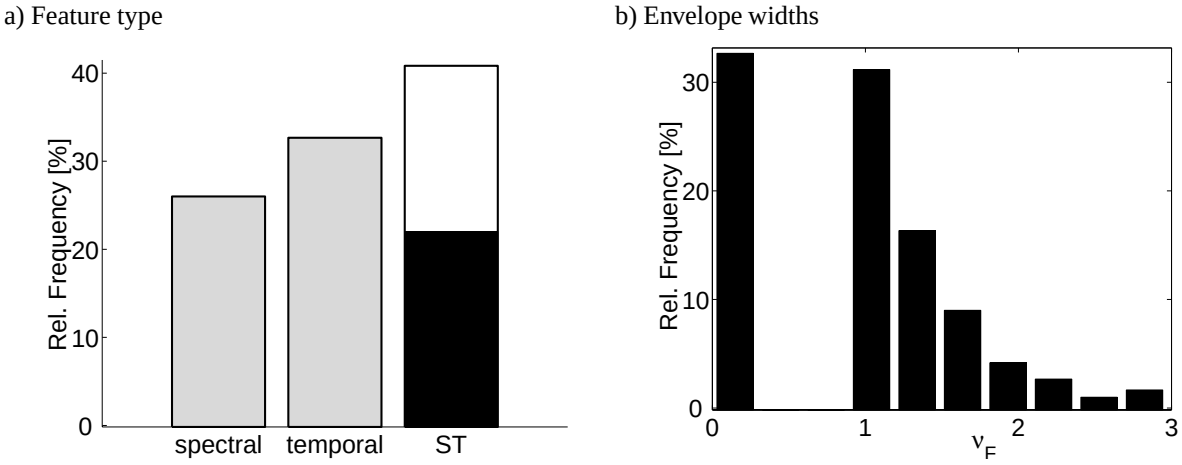


Figure 5 - Statistics for feature prototypes with Hanning envelope (total of 640 features). a: Distribution of purely spectral or temporal LSTFs (grey) and spectro-temporal filters. The latter are split in upwards (black) and downwards (white) direction, corresponding to positive or negative temporal modulation frequencies. b: Distribution of the ratio $\nu_k = T_k/2\sigma_k$.

5 Conclusions

In this paper, a number of experiments are presented that analyze the previously proposed Gabor-shaped LSTFs and investigate methods of improvement.

Regarding the number of features that should be in a set, the 30 most relevant features from a given set already show high performance. If application-specific constraints allow for more, up to 70 out of 80 should be used for optimal performance.

Another optimization problem arises with the width of the prototype envelope relative to the modulation frequency period. The wider the envelope (larger ν_x) the more selective is the filter in modulation frequency domain. However, this benefit comes with the expense of larger

prototypes, that contain more complex spectro-temporal patterns, have higher computational demand, and are not very well corresponding to physiological STRFs. In past experiments, 1.5 oscillation periods per feature ($\nu_x = 1$) were chosen ad hoc as a fixed ratio for all features in the set. The experiments in this paper indicate that halving the envelope size increases performance in clean and very high SNR conditions, while deteriorating performance for low SNR and clean training (mismatch conditions). Allowing for automatic selection of ν_x yields a distribution that peaks close to one (Fig. 5). This supports the ad hoc defined prototype. However, the overall results support a loose constraint on envelope width by means of an allowed range might be beneficial as each individual feature may have a slightly different optimal ν_x value.
Hanning-shaped LSTFs show sharper modulation frequency characteristics and lead to increased performance compared with baseline results and feature sets with Gaussian envelope.

Special thanks go to David Gelbart, Heiko Gölzer, Hynek Hermansky, Birger Kollmeier, and Nelson Morgan for their contribution of ideas. This work was supported by Deutsche Forschungsgemeinschaft (KO 942/15).

Literature

[1] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, pp. 1–15, 1997.

[2] 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.*, vol. 85, pp. 1220–1234, 2001.

[3] C. Kaernbach, "Early auditory feature coding," in *Contributions to psychological acoustics: Results of the 8th Oldenburg Symposium on Psychological Acoustics.* 2000, pp. 295–307, BIS, Universität Oldenburg.

[4] R. De-Valois and K. De-Valois, *Spatial Vison,* Oxford U.P., New York, 1990.

[5] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," in *Proc. ICSLP, Denver*, 2002.

[6] H. Hermansky, "Should recognizers have ears?," *Speech Communication*, vol. 25, pp. 3–24, 1998.

[7] M. Kleinschmidt, "Spectro-temporal Gabor features as a front end for ASR," in *Proc. Forum Acusticum, Sevilla*, 2002.

[8] T. Gramß and H. W. Strube, "Recognition of isolated words based on psychoacoustics and neurobiology," *Speech Communication*, vol. 9, pp. 35–40, 1990.

[9] T. Gramß, "Fast algorithms to find invariant features for a word recognizing neural net," in *IEEE 2nd International Conference on Artificial Neural Networks*, Bournemouth, 1991, pp. 180–184.

[10] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acustica united with acta acustica*, vol. 88, pp. 416–422, 2002.

[11] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP*, 2000.

[12] H.G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA ITRW ASR*, 2000.

[13] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, S. Kajarekar P. Jain, N. Morgan, and S. Sivadas, "QUALCOMM-ICSI-OGI features for ASR," in *Proc. ICSLP*, 2002.