EXTENDING FEATURES FOR AUTOMATIC SPEECH RECOGNITION BY MEANS OF AUDITORY MODELLING

Gero Szepannek[1], *Tamas Harczos*[2], *Frank Klefenz*[2], *Claus Weihs*[1]

[1] Department of Statistics, Dortmund University of Technology, 44221 Dortmund, Germany
[2] Fraunhofer Institute of Digital Media Technology, Ehrenbergstrasse 31, 98693 Ilmenau, Germany
szepannek@statistik.tu-dortmund.de

ABSTRACT

When investigating the benefit of auditory modelling for automatic speech recognition applications typically different features or auditory simulation models are compared. In this work the attempt of *combining* several auditory model based feature extraction schemes is pursued, as well as their further combination with standard MFCC features.

For this purpose a regularization of the common heteroscedastic discriminant analysis is introduced to summarize relevant information in feature spaces of lower dimension and uncorrelated single features.

Besides standard auditory model - based features also new features are included that rely on delay computing networks to extract relevant information from the shape of the cochlear travelling wave delay trajectory. In an empirical study statistically significant improvements are shown by combining standard MFCCs with the different features extracted from the auditory simulation model. The effect of different degrees of regularization is investigated for this task.

1. INTRODUCTION

Auditory modelling for feature extraction in ASR applications has already been investigated in a couple of papers. Some of them compare different features (e.g. [1], [2], [3]) others compare features based on different kinds of auditory modelling (e.g. [4], [5]) as opposed to standard features like MFCCs. But rarely the combination of different auditory-model based feature extraction-principles is pursued although it is probable that humans also make use of several neural information coding-principles simultaneously. The most famous coding schemes of auditory neural information transmission are *mean activity rates* at different positions along the ear as well as *phase locking* based features capturing periodic neural activity. Auditory modelling and feature extraction are explained in Section 2 and 3. Besides standard features, *delay computing networks* [6] are introduced as a tool to extract (non-frequency) information within the delay structure of the auditory neural response according to observations in [7] that the shape of the cochlear travelling wave delay trajectory also carries relevant information about the underlying sound signal.

Regularized heteroscedastic discriminant analysis (RHDA) as an extension of the common *heteroscedastic discriminant analysis* (HDA, [8]) is presented in Section 4 to combine different feature sets. The results of an empirical study are given in Section 5 and 6 and summarized in Section 7.

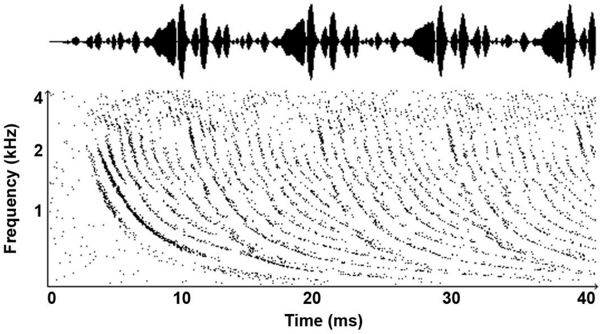

Figure 1: *Output of the auditory model for a vowel /a/.*

2. AUDITORY MODELLING

Several well-known psycho-acoustical phenomena can be traced back to sound processing in the auditory system, e.g. nonlinear frequency resolution and amplitude saturation or masking effects. An overview of the auditory processing chain is given in [9]. Basically, the sound wave is nonlinearly bandpass-filtered along the basilar membrane (BM) and transduced into electric impulses (action potentials, APs) at the auditory nerve fibres (ANFs) of different center frequency (CF) by inner hair cells (IHCs). In literature, different attempts of different degrees of biological precision can be found to imitate human auditory sound processing. For the application of this work a very detailed model of the outer ear, middle ear and BM movement including the effect of the outer hair cells (OHCs) [10] is used that is specifically designed to mimic human masking thresholds. The model is coupled to a state of the art neurophysiologically parameterized IHC/ANF model [11] resulting in the simulation of exact firing times of 251 different ANFs with CF differences of 0.1 bark along the BM. Figure 1 (bottom) shows the simulated response of the auditory simulation model to some vowel /a/. The ordinate represents the unrolled inner ear (BM) while the abscissa denotes the time. The output of the simulation model is binary and of the form

$$X_i(t) = \begin{cases} 1, & \text{AP of ANF } i \text{ at time } t \\ 0, & \text{else.} \end{cases}$$

The travelling waves of neural reaction along the cochlea are clearly recognizable. The velocity of the travelling wave is not constant (otherwise it would result in straight lines) but slows down. According to observations in [7] and [12] the shape of the cochlear delay trajectory carries information about the underlying sound signal as the wave slows down at the position of maximal BM resonance. It is also visible that different positions along the BM are differently excited (according to the signal frequencies). Furthermore, there is periodic structure in the response of the ANFs due to a phenomenon called *phase locking* (the auditory neurons tend to fire only during the positive half-waves of the signal periods). Finally, a strong response can be observed for the onset of the signal. For the studies in this paper, 50 repetitive simulations of the ANFs of different type (30 HSR, 10 MSR, 10 LSR, according to their natural distribution) are pursued. Speech signals were presented to the auditory model at 62.5 dB SPL representing a typical value for conversations.

3. FEATURE EXTRACTION

According to the above mentioned principles of neural information coding standard *place / mean rate features* (MR) and *average localized synchrony detection* (ALSD) are computed. Mean rates count the neural response at different ANFs independently of its temporal fine structure, i.e.

$$X_i^{MR} = \sum_{t \in \text{window}} X_i(t)/\text{window size}$$

A typical window size of 20 ms is used, where the speech signal can be assumed to be stationary. According to [13] groups of 8 neighbouring ANFs in the CF range of $[200, 6400]$ Hz are averaged to build a 24 dimensional feature vector. In [14] it was observed that a rather broad range of fibres is activated by signal frequencies especially at higher sound pressure levels. Thus, MR coding might not be the only way of human auditory information transmission.

ALSD features as an representatitve of temporal neural information coding are computed according to:

$$X_i^{ALSD} = \frac{1}{3}\sum_{l=i-1}^{i+1} X_l^{GSD_i} \quad \text{where} \tag{1}$$

$$X_l^{GSD_i} = A_s \tan^{-1}\left[\frac{\langle |X_l^*(t) + X_l^*(t-n_i)|\rangle - \delta}{A_s\langle |X_l^*(t) - \beta^{n_i} X_l^*(t-n_i)|\rangle}\right]$$

with $X_l^*(t)$ being the time-varying firing rate of ANF l (estimated by the post stimulus time histogram of the neural activity in time bins of $\frac{1}{14700}$ s averaged over all simulations and 8 neighbour ANFs as for X^{MR}). The $\langle . \rangle$ operator denotes temporal averaging, n_i is the period (in time bins) of the CF of ANF i. Basically, the term in the denominator checks, whether on average the neural activity is the same as it has been one (CF-)period before. The constant $\beta = 0.99$ avoids obtaining zeros in the denominator. $\delta = 60\,\text{spikes}/s \cdot dt$ corrects for spontaneous neural activity and $A_s = 4$ is a scaling constant. According to eqn. (1) the X^{ALSD} representation consists in a 22 dimensional feature vector. As phase locking decreases for high frequencies above $1-2$ kHz [15] also temporal coding can not be the only way of carrying auditory information but rather a combination of X^{MR} and X^{ALSD}.

In addition to standard MR and ALSD features also features are computed to extract information about the shape of the cochlear travelling wave: In [16], [17] and [18] features are investigated that are based on so called neural *delay computing networks* (DCNs). As such networks are less common they are briefly explained in the rest of the Section:

DCNs [6] identify curves of different shape. Figure 2 shows an exemplary 9×9 DCN. The input neurons (INs) are shown in the left vertical line. At each time t either a spike (AP) appears at an IN i (i.e. $X_i(t) = 1$) or not. At each time step $t \rightarrow t+1$, the spikes of each (vertical) layer j at delay chain (i.e. input neuron position) i are transmitted to the next layer $j+1$ with some probability $P_{ij}(t)$ or otherwise delayed for exactly one time step. The output neurons j (ONs) integrate all spikes that are in layer j at the same time t resulting in their instantaneous activity $Y_j(t)$.

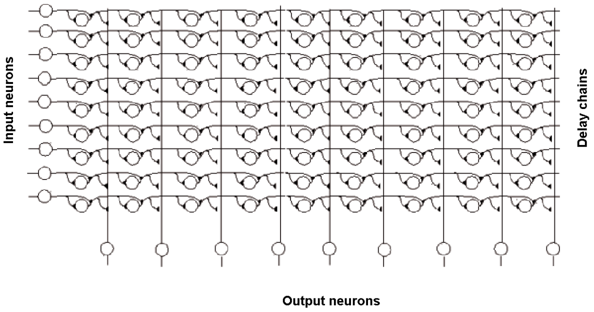

Figure 2: *Exemplary* 9×9 *delay computing network.*

During the learning process, the probabilities $P_{ij}(t)$ are trained unsupervisedly until convergence. Finally, each ON j corresponds to a specific delay structure within the response of the INs according to the trained values P_{ij}. A description of the training algorithm is given in [6] and it is shown that such networks are able to learn curves of different shape like straight lines or sinusoidals. In [19] DCNs are adapted to the output of the auditory simulation model where the ANFs (144 ANFs starting with index 21, ~ 200 Hz CF) serve as INs of the DCN. In [17] the implementation of several parallel local (PL)DCNs of smaller size is proposed where the INs of any local DCN cover only a small neigbourhood range of ANFs.

Computing mean activities $\langle Y_j(t)\rangle$ as features is not meaningful as each layer (and thus each ON) is passed by all input APs but at different times (see e.g. [18]). A specific shape of the cochlear travelling wave would lead to strong temporally concentrated activity $Y_j(t)$ of its corresponding ON j. From a neurophysiological point of view neurons do fire if their potential exceeds some threshold θ. Thus, one could argue that if some shape of the cochlea travelling wave is representative for a specific sound signal (e.g some specific phoneme) for any t of different sound the corresponding (output-)neuron should show less activity than θ, i.e. $P(\max_t Y_j(t) < \theta)$ is small for all t where the specific sound is not present. Feature representations $X_j^{DCN} = \max_t Y_j(t)$ can be motivated [17]. Note that as opposed to the formerly presented features, X_j^{DCN} can not be interpreted in a spectral manner. The only implicit frequency information depends on the CFs of the ANFs that serve as INs. Thus, PLDCNs of smaller local network size represent smaller CF ranges. After some experiments 10 parallel local networks each of size 18 are implemented (starting at ANF 21). For X^{DCN} feature extraction the values $\max_t Y_j(t)$ of any 9 neighbour output neurons are further grouped by summing up their values (this gave the best results among several investigated group sizes, for further details on parameterization of the PLDCNs see [17]). The resulting features correspond to either simultaneous or strongly delayed activity within the frequency region of a local network's input ANFs.

4. COMBINING FEATURE VECTORS

To combine several feature sets, apart from simple concatenation linear transformations of the combined data vectors have become popular to both reduce the dimensionality (and thus the number of free model parameters) as well as to decorrelate the resulting feature vectors (allowing to use diagonal covariance matrices for HMM back end modelling).
Standard methods for this task are principal component (PCA) transformation as well as cepstral transformation. Both methods are known to produce uncorrelated features. Principal component transform returns uncorrelated features w.r.t. some common mean (of all recognizable classes). To be able to perform recognition, the conditional feature distributions given the classes is supposed to differ. A well known linear dimensionality reduction transform that maximizes the average distances of the class means in the transformed space w.r.t. their (common) covariance is linear discriminant analysis (LDA). An extension of LDA that also handles different covariance structures of the class distributions is heteroscedastic discriminant analysis (HDA, [8]), where a transformation matrix $\mathbf{A}$ is determined that maximizes the likelihood (in the transformed space) $\mathbf{y} = \mathbf{A}'\mathbf{x}$ under the assumption of normality. For the class specific means and covariances it is further assumed

$$\mu_{\mathbf{k}} = \begin{bmatrix} \mu_{\mathbf{k}}^{\mathbf{q}} \\ \mu_{\mathbf{0}}^{\mathbf{p-q}} \end{bmatrix} \text{ and}$$

$$\mathbf{\Sigma_k} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{k}}^{\mathbf{q}} & 0 \\ 0 & \mathbf{\Sigma}_{\mathbf{0}}^{\mathbf{p-q}} \end{pmatrix},$$

i.e. the class distributions only differ in the first q components of the transformed space (where $p > q$ is the dimension of the original space and k denotes the class). Finally, only the first q components of $\mathbf{y}$ are used for back end estimation. An efficient algorithm under the assumption of diagonality in the transformed space is given in [20]. The estimates for class means and covariances in the transformed space are given by

$$\begin{aligned} \hat{\mu}_{\mathbf{k}}^{\mathbf{q}} &= (\mathbf{A^p})'\bar{\mathbf{x}}_{\mathbf{k}}, \\ \hat{\mu}_{\mathbf{0}}^{\mathbf{p-q}} &= (\mathbf{A^{p-q}})'\bar{\mathbf{x}}, \\ \hat{\mathbf{\Sigma}}_{\mathbf{k}}^{\mathbf{q}} &= \mathrm{diag}((\mathbf{A^q})'\mathbf{W_k}(\mathbf{A^q})) \text{ and} \\ \hat{\mathbf{\Sigma}}_{\mathbf{0}}^{\mathbf{p-q}} &= \mathrm{diag}((\mathbf{A^{p-q}})'\mathbf{T}(\mathbf{A^{p-q}})) \end{aligned}$$

(see [8]) where the upper indices ^{q} (resp. $^{p-q}$) denote the first q (last $p-q$) components and $\bar{\mathbf{x}}_{\mathbf{k}}$ and $\mathbf{W_k}$ (resp. $\bar{\mathbf{x}}$ and $\mathbf{T}$) are the class specific (resp. total) mean and covariance matrix estimations of the features in the original space. In HDA, besides the enlarged model flexibility the number of free model parameters increases and thus finding the transformation becomes less stable. Therefore, in [21] regularization is proposed where several components of $\mathbf{A}$ are fixed in advance to be 0 (e.g. some block structure of $\mathbf{A}$ is required). This is meaningful if the dimensionality reduction is desired to be performed separately on several subgroups of variables (e.g. the original features on one hand and their first and second order Δ derivatives on the other hand in [21]). By construction, this kind of regularization is not meaningful for feature combination. In [20] stabilization is proposed by smoothing the estimate (in the original space)

$$\mathbf{W}_{\mathbf{k}}^{\mathbf{smoothed}}(\lambda) = \lambda\mathbf{W_k} + (1-\lambda)\mathbf{W_{pooled}}$$

where $\mathbf{W_{pooled}}$ is the common pooled equal covariance estimate for all classes and $\lambda \in [0,1]$.
Friedman [22] proposes for classification tasks:

$$\mathbf{W}_{\mathbf{k}}^{\mathbf{Fried}}(\lambda) = \frac{(\lambda)N_k\mathbf{W_k} + (1-\lambda)N\mathbf{W_{pooled}}}{(\lambda)N_k + (1-\lambda)N}.$$

that takes into account that covariance estimates of smaller class sizes N_k are less stable. He furthermore proposes an additional shrinkage of $\mathbf{W}_{\mathbf{k}}^{\mathbf{Fried}}$ towards diagonality:

$$\mathbf{W}_{\mathbf{k}}^{\mathbf{RDA}}(\lambda,\gamma) = \gamma\mathbf{W}_{\mathbf{k}}^{\mathbf{Fried}}(\lambda) + (1-\gamma)\frac{tr(\mathbf{W}_{\mathbf{k}}^{\mathbf{Fried}}(\lambda))\mathbf{I}}{p}$$

where $\mathbf{I}$ is the identity and $(\lambda,\gamma) \in [0,1]^2$. This idea goes back to [23]. Let *regularized heteroscedastic discriminant analysis (RHDA)* be the transform that is obtained by HDA with covariance estimates $\mathbf{W}_{\mathbf{k}}^{\mathbf{RDA}}(\lambda,\gamma)$ in the original space. Note that the extremes $(\lambda,\gamma) = (1,1)$ and $(\lambda,\gamma) = (0,1)$ represent HDA and LDA. In [8] it is shown that the LDA transform also maximizes the likelihood as in HDA under equal covariance assumption of the classes. For this work RHDA transforms are evaluated on parameter grids $(\lambda,\gamma) \in \{0, 0.25, 0.5, 0.75, 1\}^2$ and different choices of q. The classes for training are chosen to be phoneme states of an initial alignment using standard MFCC based recognizers according to the results of [24] (see Section 5) in order to provide a unified basis for all feature vectors under investigation. According to [25] $X_\Delta(t) := X(t) - X(t-20ms)$ and $X_{\Delta\Delta}(t)$ are explicitly computed before transformation.

5. EVALUATION STUDY

An evaluation study is pursued on the TIMIT data base [26]. The core set is used as proposed by the developers, consisting of training (test) sets of 576 (192) utterances (SI and SX).

- Left-to-right HMMs are trained for each of the 61 monophones, consisting of 3 states each.
- The feature distributions given the states are modelled as gaussian mixtures.
- State initialization is done by linearly segmenting the utterances.
- After any three iterations of Baum-Welch reestimation gaussians are split into mixtures of 2 gaussians until there were finally 8 mixtures of gaussians per state.
- Bigrams are used as grammar models.
- Prediction is done by the Viterbi algorithm.

The implemented back end corresponds to observed results for different parameterizations described in [27]. Even if they are observed for non-auditory feature sets it appears meaningful to use them as an initial guess. Concerning the number of mixtures one may suppose that it results from several groups of subpopulations within the data that exist independently from the explicit feature choice of features. Some brief attempts have been made to change the number of mixtures or to implement generalized triphone models, but without an increase in recognition rate. The latter is supposed to result from the limited size of training data having a strong effect on the performance [27]. As the focus of the study was to investigate the benefit from feature combination of different neural information coding schemes no further attention was invested into optimizing the HMMs by additional tying methods. Implementation is done using the HTK toolkit [28].

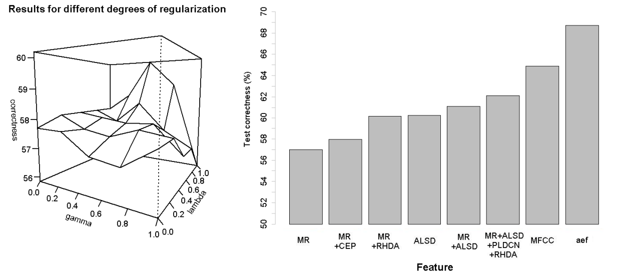

Figure 3: *Recognition rates for different degrees of regularization (left) and different feature sets (right).*

According to Lee and Hon [29] for recognition 39 phone classes are defined from the original 61 phones. Besides comparing the recognition correctness on the test data statistical tests are implemented to investigate significance of the difference of the observed results. 5×2 fold cross-validation is used for testing as it is proposed in [30] with additional Bonferroni corrections for multiple testing [31].

6. RESULTS AND DISCUSSION

A first investigation concerned dimensionality reduction methods evaluated on standard *mean rate* features. Using RHDA shows the strongest improvements (60.18%) as opposed to HDA (55.83%), LDA (57.60%) and cepstral transformation (57.97%). Figure 3 (left part) shows the results for different parameters (λ, γ) and $\dim(\mathbf{y}) = 21$. Consistently optimal results (for different dimensionalities and original feature sets) are obtained for parameter choices $(\lambda, \gamma) \sim (1, 0.5)$. This underlines the importance of modelling class specific covariance matrices. Box M test (see e.g. [32], exemplarily performed on different subsamples of variables and classes) either strongly refuses the hypothesis of equal covariance matrices of the classes or runs into matrix inversion difficulties due to nearly collinearity of the covariance matrix estimation in the original space. The latter leads to high variance of the estimates ([23]) and might explain the benefit of regularization as it is done by a parameter $\gamma < 1$. Figure 4 shows an example of the covariance matrix estimates for two different classes in the original feature space. Class specific covariance matrices are clearly recognizable. Due to their nature components with similar indices show strong correlations.

Figure 3 (right part) shows the performance of ASR systems based on different feature sets: the combination of X^{MR} (56.98%) and X^{ALSD} (60.25%) significantly improves the recognition performance (61.11%) by means of the simple convex combination ($X_k^{new} = 0.15 \cdot X_k^{MR} + 0.85 \cdot X_k^{ALSD}$) of the standardized original features. Adding PLDCN features (53.35%) leads to further statistically significant improvements (62.13%). The small recognition performance of PLDCNs alone can be explained by the fact that relevant frequency information has been dropped away in this case.

Standard hamming windowed MFCC features result in 64.90% correctness (which corresponds to the results in [27] for similar amounts of training data, similarly, using PLPs resulted in 64.27%). This good performance of MFCCs alone might be explained as they already imitate several phenomena of auditory sound processing like mel-scaled frequency transformation or logarithmic amplitude saturation [4].

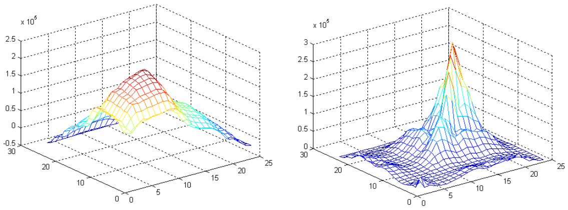

Figure 4: X^{MR} *Covariance matrices for different classes.*

Finally, performing RHDA on the combined MFCC + MR + ALSD + PLCDN (auditory extended feature, **aef**) vector strongly improved the recognition rates up to 68.72% This result shows the benefit of combining different auditory and non-auditory features and underlines the hypothesis that the auditory neural response contains additional relevant information for automatic speech recognition. Statistical tests underline significance of difference in recognition performance of four compared pairs of feature sets under investigation:

feature set 1	feature set 2	p-value
MR	MR (RHDA)	0.0067
MR	MR/ALSD	0.0046
MR/ALSD	MR/ALSD/PLDCN (RHDA)	0.0429
MFCC	**a**uditory **e**xtended **f**eature set	$2.11 \cdot 10^{-5}$

Table 1: p values for tests on equal performance.

7. SUMMARY AND OUTLOOK

Different principles of feature extraction from a detailed neurophysiologically parameterized auditory simulation model are compared in this paper. Furthermore, motivated by the fact that humans are probable to make use of several ways of auditory neural information encoding, the combination of different auditory model based feature sets is investigated. For this purpose the method of *regularized heteroscedastic discriminant analysis* is introduced to condense relevant information of several different feature sets in an uncorrelated vector of lower dimension. Especially the aspect of regularization showed to be beneficial while at the same time allowing to model heteroscedastic covariance matrices of different classes. An increased recognition performance has been observed for the combination of different auditory model based information encoding schemes (i.e. place/mean rates, phase locking, delay computing). Moreover, it has been shown that extended recognizers based on combined MFCC / auditory model based features can significantly improve the recognition rate compared to the simple use of the common MFCCs. The results lead to the conclusion that there is additional relevant information included within the auditory model based features. For further studies emphasis should be laid on back end optimizing by implementing tying approaches on larger data bases. Furthermore, the behavior of the combined feature sets under adverse conditions like noise or reverberation should be investigated.

Acknowledgement: This work has been supported by the German research foundation (DFG).

REFERENCES

[1] C. Jankowski and R. Lippman "A comparison of signal processing front ends for automatic speech recognition", *Speech and Audio Proc.* vol. 3(4), pp. 286-293, 1995.

[2] M. Holmberg, D. Gelbart and W. Hemmert, "Speech encoding in a model of peripheral auditory processing: quantitative assessment by means of automatic speech recognition", *Speech Communication* vol. 49, pp. 917-932, 2007.

[3] A. Ali, J. van der Spiegel and P. Muller, "Robust auditory-based speech processing using the ALSD", *IEEE Trans. Speech and Audio Proc.*, vol. 10(5), pp. 279–295, 2002.

[4] F. Perdigao and L. Sa, "Auditory models as front-ends for speech recognition", *Proc. NATO ASI on Computational Hearing*, 1998.

[5] M. Holmberg, D. Gelbart and W. Hemmert, "Auditory based automatic speech recognition", *Proc. SAPA 2004*.

[6] A. Brückmann, F., Klefenz and A. Wünsche, "A neural net for 2d-slope and sinusoidal shape detection", *Int. Scient. J. Computing*, vol. 3(1), pp. 21–26, 2004.

[7] S. Greenberg, "The significance of the cochlear travelling wave for theories of frequency analysis and pitch", In: Lewis, E. and Steele, C. and Lyon, R. (eds): *Diversity in Auditory Mechanics*, World Sc. Publ., 1997.

[8] N. Kumar, *Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition.* PhD Thesis, Johns Hopkins University, Baltimore, MD, 1997.

[9] G. Szepannek, F. Klefenz and C. Weihs, "Schallanalyse - Neuronale Repräsentation des Hörvorgangs als Basis", *Informatik Spektrum*, vol. 28(5), pp. 289–295, 2005.

[10] F. Baumgarte, *Ein physiologisches Gehörmodell zur Nachbildung von Wahrnehmungsschwellen für die Audiokodierung*. PhD Th., Univ. Hannover, D, 2000.

[11] C. Sumner, E. Lopez-Poveda and R. Meddis, "A revised model of the inner hair cell and auditory-nerve complex", *JASA*, vol. 111(5), pp. 2178–2188, 2002.

[12] B. Moore, "Coding of sounds in the auditory system and its relevance to signal processing and coding cochlear implants", *Otology and Neurotology*, vol. 24, pp. 243–254, 2003.

[13] J. Allen, "How do humans process and recognize speech?", *IEEE TSAP*, vol. 2(4), pp.567–577, 1994.

[14] M. Sachs and E. Young, "Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate", *JASA* vol. 66(2), pp. 470–479, 1979.

[15] D. Johnson, "The relationship between spike rate and synchrony in responses of auditory-nerve fibres to single tones", *JASA* vol. 68(4), pp. 1115-1122, 1980.

[16] T. Harczos, G., Szepannek, A., Katai and F. Klefenz, "Auditory model based vowel classification", *Proc. IEEE Biomedical Circuits and Systems Conference 2006*, pp.69–72.

[17] G. Szepannek and C. Weihs, "Explorative development of information extraction schemes for speech recognition from simulated auditory neural response data via parallel local Hubel-Wiesel networks", *Research paper 2/2006*, Dept.Statistics, Dortmund.

[18] G. Szepannek, T. Harczos, F. Klefenz, A. Katai, P. Schikowski and C. Weihs, "Vowel classification by a perceptually motivated neurophysiologically parameterized auditory model", In: Decker, R. and Lenz, H.: *Advances in Data Analysis*, Springer, pp. 653–660.

[19] T. Harczos, F. Klefenz and A. Katai, "A neurobiologically inspired vowel recognizer using Hough-transform – a novel approach to auditory image processing", *Proc. of Visapp 2006*, vol.1: pp. 251-256.

[20] L. Burget, "Combination of speech features using smoothed heteroscedastic linear discriminant analysis", *Proc. Interspeech 2004, Jeju, Korea*, pp. 2549–2552.

[21] H. Erdogan, "Regularizing heteroscedastic discriminant analysis for speech recognition", *Proc.* 13*th* *IEEE Sig. Proc. and Comm. Appl. Conf. 2005*, pp. 110-107.

[22] J. Friedman, "Regularized discriminant analysis", *J. American Statistical Asscociation* vol. 84, pp. 165-175, 1989.

[23] P. Di Pillo, "The application of bias to discriminant analysis", *Communications in Statistics - Theory and Methods*, vol. 5(9), pp. 843-854, 1976.

[24] J. Duchateau, K. Demunynck, D. Compernolle and P. Wambacq, "Class definition in discriminant feature analysis", *Proc. Eurospeech 2001, Geneva, Switzerland vol. 2)*, p. 1621-1624.

[25] T. Eisele, R. Haeb-Umbach, and D. Langmann, "A comparative study of linear feature transformation techniques for automatic speech recognition", *Proc. ICSLP96, Philadelphia*, vol. 1, pp. 252-255.

[26] J. Garofolo, L. Lamel, W. Fiesher, J. Fiscus, D. Palett and N. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus", *Tech. Rep. NISTIR 4930*, NIST, Gaithersburgh, MD, 1993.

[27] E. Gouws, K. Woolvaardt, N. Kleynhans and E. Barnard, "Apropriate baseline values for HMM-based speech recognition", `http://www.meraka.org.za/pubs/gouwse04HMMbaselines.pdf`, Univ. Pretoria, South Africa, 2004.

[28] S. Young, G. Everman, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The HTK-book (v 3.3)*, Cambridge University Engineering Dept., 2005.

[29] K. Lee and F. Hon, "Speaker-independent phone recognition using hidden Markov models", *IEEE Trans. on Speech and Signal Proc.*, vol. 37(11), pp. 1641–1648, 1989.

[30] T. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms", *Neural Computation*, vol. 10(7), pp. 1895-1923, 1998.

[31] S. Holm, "A simple sequentially rejective multiple test proceedure", *Scand. J. Statistics*, vol. 6, pp. 65-70, 1976.

[32] K. Mardia, J. Kent and J. Bibby, *Multivariate Analysis*. Academic Press, London, 1979.