Spectro-Temporal Features with Distribution Equalization

Samuel K. Ngouoko M.[1,2], *Martin Heckmann*[2], *Britta Wrede*[1]

[1] Research Institute for Cognition and Robotics; Bielefeld University, D-33615 Bielefeld, Germany.
[2] Honda Research Institute GmbH, D-63073 Offenbach/Main, Germany.
(sngouoko, bwrede)@cor-lab.uni-bielefeld.de, martin.heckmann@honda-ri.de

Abstract

We could show in the past that Hierarchical Spectro-Temporal (HIST) features improve the performance of Automatic Recognition Systems (ARS) of speech in difficult environments when they are combined with conventional speech spectral features. The target here is to improve the noise robustness of the HIST features by investigating a channel distribution equalization in our feature hierarchy. Thereby, we determine the empirical cumulative distribution of the speech training data set, which is referred to as reference distribution. Afterwards, a distribution adjustment of the training as well as test data is performed with respect to the reference distribution. We carry out the above mentioned distribution equalization in the preprocessing step as well as after each feature extraction step of our HIST feature extraction framework. We evaluate the benefits of such an equalization in the HIST feature extraction process with different noise types.
Index Terms: Spectro-temporal features, distribution equalization.

1. Introduction

In severe acoustical environments, e.g. when the noise exhibits nonstationary characteristics, the performance of Automatic Speech Recognition (ASR) systems decreases remarkably, especially in comparison to humans [1].

Common spectral speech features as the Mel Frequency Cepstral Coefficients (MFCCs) or RelAtive SpectrAl (RASTA) features [2] show good performance in clean conditions but strongly deteriorate in the presence of noise. In order to enhance the feature representation and consequently improve their performance in difficult environments several normalization methods in the feature space have been proposed. On one hand there is the *Cepstral Mean Subtraction* (CMS), where the global shift of the cepstrum is reset to zero, on the other hand the *Mean Variance Normalization* (MVN), which is an extension of the CMS in which, additionally to the removal of the mean of each feature vector, their standard deviation is normalized to unity [3, 4]. These methods improve the performance of MFCCs in difficult environments, but their performance decreases (not significantly) in clean condition. Furthermore, a better noise robustness of the features is also achieved by performing a Distribution Equalization (DEQ), where the nonlinear distortions caused by noise are compensated. Hereby, the distribution of the features is adjusted such that it becomes similar to a reference distribution often chosen as a normal distribution.

Spectro-temporal features gave promising results in severe environments. These features are inspired by neurophysiological findings and allow to capture the joint spectro-temporal dynamics of speech. Unlike standard features, they are able to detect diagonal structures in the spectro-temporal representation as formant transitions. Most of them use Gabor filters [5, 6, 7, 8, 9]. Alternatively, we developed features inspired by a hierarchical system for visual object recognition [10]. The feature extraction is organized in two hierarchical layers and we refer to them as Hierarchical Spectro-Temporal (HIST) features [11, 12]. An overview of the feature extraction scheme is depicted in Fig. 1. These spectro-temporal features are more robust in difficult environments compared to conventional speech features and especially in combination with them.

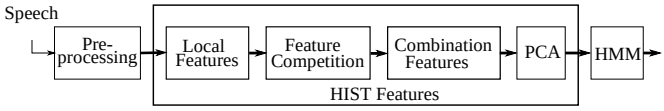


Figure 1: Overview of the feature extraction process [12].

Since the normalization procedures mentioned above contribute to improve the performance of spectral features, we aim at applying these methods in our spectro-temporal feature extraction process. MVN has already been shown to perform well for Gabor based spectro-temporal features [13]. In this paper we investigate the influence of DEQ in our hierarchical feature extraction framework in the different processing steps.

This contribution is organized as follows. Section 2 reviews brievly the main steps of the HIST feature extraction. Section 3 is devoted to the description of the feature normalization methods. The simulation results, a conclusion and discussion constitute the last sections, respectively.

2. HIST feature extraction

The HIST feature extraction process consists of several steps [11, 12]:

- In the first step the input speech signals are transformed into the spectro-temporal domain using a *Gammatone*

filterbank [14], which models the peripheral processing done by the cochlea in the auditory system. The filterbank has 128 channels for a frequency range from 80 Hz to 8 kHz. From this we obtain spectrograms by rectifying and low-pass filtering of the filterbank response. The sampling rate is then reduced to 400 Hz. Afterwards, the spectral components of the excitation and radiation are suppressed by amplifying the frequency magnitude by +6 dB/oct. This is termed *preemphasis*. After preemphasis a spectral filtering is carried out by smoothing the spectrogram in the frequency direction with channel-dependent Difference-of-Gaussian (DoG) operators. This contributes to the suppression of the harmonics in the spectrogram. Hence, the formant structure is enhanced. Finally, we approximated the loudness perception using the 15th-root non-linear function [15].

- Afterwards, the local features are determined as the absolute of the 2D convolution of the input spectrogram with a set of receptive fields learned with Independent Component Analysis (ICA) [16]. It follows a competition of coequal features using the Winner Take Most (WTM) algorithm for removing less active neurons and improving the feature selectivity. This builds the first layer of our hierarchy.

- In the second layer complex features are obtained by combining local features over a large frequency range. The features are learned using Non-Negative Sparse Coding[17].

- Finally, the features from the second layer are decorrelated using the Principal Component Analysis (PCA).

3. Feature normalization procedures

Several techniques are commonly used for the enhancement of the feature representation and have contributed to increase the robustness of speech recognition in difficult environments [4]. In this paper we considered the MVN and the DEQ techniques [3].

3.1. Mean and Variance Normalization

In the MVN technique, the speech signal x_0 is normalized according to the equation (1) such that its mean μ_0 and its variance σ_0^2 are changed to 0 and 1, respectively.

$$x_1 = F(x_0) = \frac{x_0 - \mu_0}{\sigma_0}, \tag{1}$$

where x_1 is the normalized speech signal [3, 4].

3.2. Distribution Equalization

Let x_0 be a variable following a probability density function (pdf) $p_0(x_0)$. The goal of this technique is to define a transformation $x_1 = F(x_0)$ that converts the pdf $p_0(x_0)$ into the reference pdf $p_1(x_1) = p_{\text{ref}}(x_1)$ according to the expression [3]:

$$p_1(x_1) = p_0(G(x_1))\frac{\partial G(x_1)}{\partial x_1}, \tag{2}$$

where $G(x_1)$ is the inverse transformation of $F(x_0)$. The relationship between the cumulative probabilities associated with these probability distributions is given by

$$\begin{aligned} C_0(x_0) &= \int_{-\infty}^{x_0} p_0(x_0')dx_0' \\ &= \int_{-\infty}^{F(x_0)} p_0(G(x_1'))\frac{\partial G(x_1')}{\partial x_1'}dx_1' \\ &= \int_{-\infty}^{F(x_0)} p_1(x_1')dx_1' \\ &= C_1(F(x_0)) \end{aligned} \tag{3}$$

and therefore, the desired transformation $x_1 = F(x_0)$ is obtained from (3) as

$$x_1 = F(x_0) = C_1^{-1}[C_0(x_0)] = C_{\text{ref}}^{-1}[C_0(x_0)], \tag{4}$$

where $C_0(x_0)$ is the cumulative distribution of the speech signal and C_{ref}^{-1} is the inverse cumulative probabilty function of the reference distribution. The cumulative distribution function (CDF) is determined as the cumulative sum of the feature sample frequency obtained from the histogram normalized by the maximal feature sample frequency. Since the cumulative distribution is a monotonically increasing function, we performed the inverse CDF by finding for each CDF value of the current feature the corresponding feature sample value from the reference distribution. We considered 500 intervals between $\mu_i - 4\sigma_i$ and $\mu_i + 4\sigma_i$ for determining the cumulative distribution, where μ_i and σ_i represent the mean and standard deviation of the *i*-th channel or feature vector component. Instead of considering the normal distribution $\mathcal{N}(0, 1)$ as reference distribution, we estimated the reference distribution based on a reduced training data set in clean conditions.

The distribution equalization is characterized by the fact that nonlinear distortions are compensated, although its efficiency strongly depends on the quality of the estimated cumulative distribution. Additionally, the distribution equalization is supposed to correct monotonic transformations, which can lead to a loss of information since the noise renders the current transformation nonmonotonic.

4. Experimental results

For the evaluation we use TIDigits [18], a database for speaker independent continuous digit recognition. We corrupted the data with different types of noise (white, factory, babble and car) from the Noisex database [19] at Signal-to-Noise Ratio (SNR) levels from -5dB $\ldots$ inf (clean signal). The recognition is performed with Hidden Markov Models (HMMs) trained on clean signals with HTK [20] using, as defined in the Aurora-2 experimental framework [21], whole

	white	factory	babble	car
RASTA-PLP$_{baseline}$	43.1	41.2	35.1	19.7
HIST$_{baseline}$	33.4	32.0	81.9	20.4
R+H$_{baseline}$	24.4	27.1	71.9	15.4
RASTA-PLP$_{MVN}$	26.4	26.8	30.9	11.6
HIST$_{MVN}$	34.3	40.4	64.1	20.8
R+H$_{MVN}$	23.2	30.1	54.0	11.2

Table 1: Average word error rates in % for the different feature types when the specified noise types at SNR values ranging from $-5\,\mathrm{dB}\ldots$ inf were added and the MVN was applied.

word HMMs containing 16 states without skip transitions and a mixture of 3 Gaussians with a diagonal covariance matrix per state. The features were also learned with clean signals and the speakers in the training set differ from those in the test set. The normalization or equalization procedures are applied during both training and recognition. We use a combination of HIST and RASTA-PLP features as we obtained previously good recognition scores with such a combination, and we consider RASTA-PLP features as baseline since they exhibit superior performance compared to MFCCfeatures [12]. In the tables and figures below, the abbreviation R+H is referred to as RASTA-PLP+ HIST.

For determining the mean, standard deviation as well as the cumulative distribution in each noise situation, we used a reduced data set of 120 utterances consisting of all types of speakers (woman, man, boy and girl). First, we applied the MVN in the feature space considering the RASTA-PLP features with delta and delta coefficients, the HIST features with delta and delta coefficients after PCA and the combination of them. Tables 1, 2 show the average recogniton performance and the average relative performance improvement for each type of noise, respectively. Hereby, the average relative improvement is calculated as the average of the relative improvements for each SNR level from $-5\,\mathrm{dB}\ \ldots$ inf. Features without normalization or equalization are referred to as *baseline*. The results demonstrate that the performance of RASTA-PLP with MVN compared to RASTA-PLP$_{baseline}$ is improved for each noise type. We notice a performance improvement of about 15% for babble noise and of about 55% or more for the other ones. We also observe that the RASTA-PLP$_{MVN}$ performance is superior to HIST features with MVN. However, by combining both HIST and RASTA-PLP features we observe a performance increase for white, factory and car noise by about 16% on average. Therefore, we assess that the complementarity of both feature information contributes to achieve better recognition scores. Nevertheless, RASTA-PLP performance still remains the best for babble noise.

Next, we applied the distribution equalization on RASTA-PLP features and also in our HIST feature extraction process, e.g. after the preprocessing (preproc.), after extraction of local (loc.) and complex (comp.) features, and after the PCA (PCA). Tables 3, 4 show the average recogniton performance

	white	factory	babble	car
RASTA-PLP$_{baseline}$	-56.0	-73.5	-14.6	-52.0
HIST$_{MVN}$	-51.0	-95.9	-318.0	-105.3
R+H$_{MVN}$	24.5	6.5	-171.2	17.6

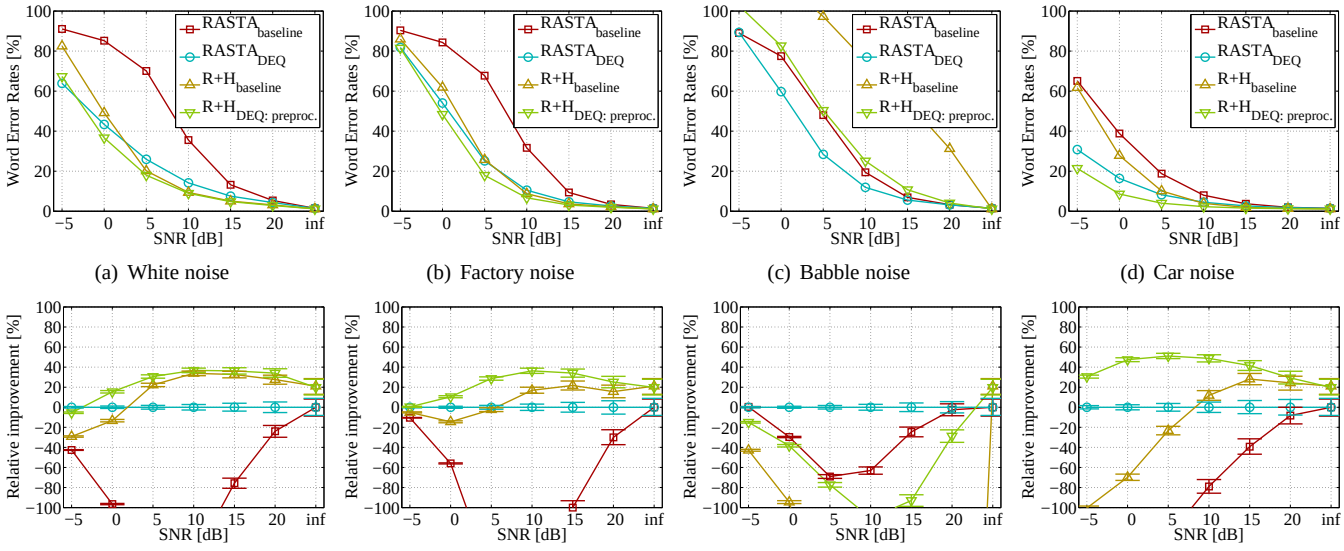
Table 2: Average relative improvement in % w.r.t RASTA-PLP$_{MVN}$ for the different feature types when the specified noise types at SNR values ranging from $-5\,\mathrm{dB}\ldots$ inf were added and the MVN was applied.

	white	factory	babble	car
RASTA-PLP$_{DEQ}$	23.0	25.8	28.5	9.4
HIST$_{preproc.}$	26.9	27.4	47.5	9.2
HIST$_{loc.}$	47.9	51.2	70.0	30.2
HIST$_{comp.}$	73.7	70.2	83.3	64.2
HIST$_{PCA}$	31.3	36.6	57.6	18.6
R+H$_{preproc}$	20.0	23.0	39.6	5.8
R+H$_{loc.}$	31.9	36.9	54.0	20.2
R+H$_{comp.}$	60.5	58.6	74.3	49.2
R+H$_{PCA}$	21.1	27.4	44.9	9.5

Table 3: Average word error rates in % for the different feature types when the specified noise types at SNR values ranging from $-5\,\mathrm{dB}\ldots$ inf were added and the DEQ was applied.

and the relative performance improvement for each type of noise, respectively. The obtained results show that applying the DEQ a performance improvement for all (or almost all) noise types is achieved as well for RASTA-PLP features as for HIST features after the preprocessing and the PCA. In contrast, we observe a performance deterioration of the HIST features when the equalization is carried out in the intermediate layers e.g. after the local and complex features. Furthermore, we notice that the RASTA-PLP$_{DEQ}$ features outperform the HIST features with DEQ in general. However, by combining the HIST and RASTA-PLP features we notice a considerable improvement of about 25% on average compared to RASTA-PLP$_{DEQ}$ for all noise types, except for babble noise. The best performance is achieved by combining the RASTA-PLP$_{DEQ}$ with the HIST$_{preproc.}$ features. Further, we compared this best performance with the corresponding baseline (R+H$_{baseline}$) in Table 4. This comparison is also illustrated in Fig. 2. A relative average improvement of about 23% is obtained for white and factory noise, and an improvement of about 40% for car noise. However, the performance of RASTA-PLP$_{DEQ}$ alone is still superior to the combination of features for babble noise. In contrast, when applying the equalization after the local or combinations features, the performance of the combination increases compared to HIST only, but still remains inferior to the performance of the RASTA-PLP features. Therefore, a distribution equalization in the intermediate layers of the HIST feature extraction scheme seems not to be beneficial.

We also performed experiments, where we estimated the mean, variance and cumulative distribution for each utterance individually and applied them on the same utterance. Tables



(a) White noise (b) Factory noise (c) Babble noise (d) Car noise

(e) White noise (f) Factory noise (g) Babble noise (h) Car Noise

Figure 2: Word error rates (a, b, c, d) and relative improvement w.r.t RASTA-PLP with DEQ (e, f, g, h) of the features when the cumulative distribution was estimated only from reduced data set.

	white	factory	babble	car
RASTA-PLP$_{baseline}$	-79.9	-81.0	-26.9	-71.9
HIST$_{preproc.}$	-41.6	-52.5	-146.5	-51.7
HIST$_{loc.}$	-126.5	-169.0	-369.6	-232.9
HIST$_{comp.}$	-463.7	-559.7	-783.3	-881.8
HIST$_{PCA}$	-57.7	-83.4	-244.1	-110.6
R+H$_{preproc.}$	23.9	22.1	-49.0	38.3
R+H$_{loc.}$	-13.9	-43.6	-221.8	-69.2
R+H$_{comp.}$	-284.5	-360.4	-568.7	-554.9
R+H$_{PCA}$	20.6	10.1	-100.2	15.9

Table 4: Average relative improvement in % w.r.t RASTA-PLP with DEQ for the different feature types when the specified noise types at SNR values ranging from $-5\,\mathrm{dB}$ ... inf were added and the DEQ was applied.

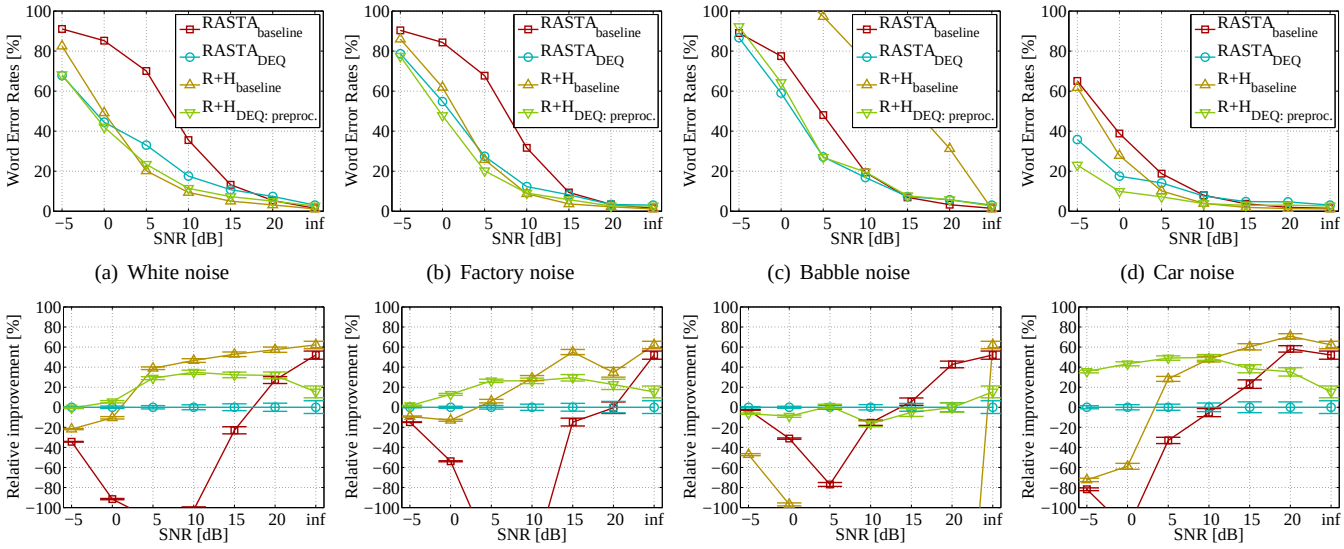
	white	factory	babble	car
RASTA-PLP$_{DEQ}$	26.3	26.8	29.4	12.5
HIST$_{preproc.}$	29.0	29.7	48.0	12.8
HIST$_{loc.}$	43.6	53.2	68.5	26.2
HIST$_{comp.}$	67.6	68.9	82.9	60.2
HIST$_{PCA}$	35.0	38.9	59.4	20.3
R+H$_{preproc}$	22.9	23.6	31.3	7.5
R+H$_{loc.}$	26.7	34.3	48.6	14.5
R+H$_{comp.}$	60.6	61.2	79.7	44.9
R+H$_{PCA}$	22.6	28.2	43.1	11.4

Table 5: Average word error rates in % for the different feature types when the specified noise types at SNR values ranging from $-5\,\mathrm{dB}$ ... inf were added and the DEQ was applied. The cumulative distributions were estimated using each utterance and applied on the corresponding utterance.

5, 6 show the corresponding average recogniton performance and the relative performance improvement for each type of noise, respectively. By comparing the results in Tables 5, 6 with those in Tables 3, 4 respectively, we remark that the feature performance behaviors are similar, when the statistics were estimated on each utterance as well as on the reduced data set. Hereby, the best performance is also achieved by combining the RASTA-PLP$_{DEQ}$ with the HIST$_{preproc.}$ features. Fig. 3 depicts the comparison between this best performance and the corresponding baseline. We can see that the performance of R+H$_{baseline}$ for Signal-to-Noise Ratio (SNR) greater than $10\,\mathrm{dB}$ is superior and vice-versa for white, factory and car noises. In the case of babble noise R+H$_{preproc.}$ outperforms the baseline and is roughly equal to RASTA-PLP$_{DEQ}$.

5. Conclusion

In this contribution we investigated a MVN as well as distribution equalization on HIST and RASTA-PLP features. The MVN has been performed only on the final features, while the DEQ was performed after each HIST feature extraction step. Two cases for estimating the statistics have been used: on one hand on a reduced data set containing 120 utterances, on the other hand on each utterance. The simulation results showed that, as expected, the performance of RASTA-PLP features improved a lot from the DEQ when a larger number of utterances (in our case 120) were used to estimate the statistics but also when only the current utterance was used. On the other hand the HIST features benefited also from DEQ for the case when applying it after the preprocessing or at the final stage (after PCA). Yet for the intermediate steps the performance decreased. When combining HIST and RASTA-PLP features after applying DEQ we saw also a notable gain in



(a) White noise (b) Factory noise (c) Babble noise (d) Car noise

(e) White noise (f) Factory noise (g) Babble noise (h) Car Noise

Figure 3: Word error rates (a, b, c, d) and relative improvement w.r.t RASTA-PLP with DEQ (e, f, g, h) of the features when the cumulative distribution was estimated only from the current utterance.

	white	factory	babble	car
RASTA-PLP$_{baseline}$	-40.6	-48.0	-3.7	-15.5
HIST$_{preproc.}$	-51.3	-75.4	-140.2	-60.2
HIST$_{loc.}$	-102.4	-190.2	-255.7	-129.9
HIST$_{comp.}$	-251.7	-362.7	-472.8	-455.3
HIST$_{PCA}$	-47.9	-78.2	-182.1	-66.0
R+H$_{preproc.}$	21.2	19.2	-3.0	38.1
R+H$_{loc.}$	3.0	-26.8	-86.9	-2.5
R+H$_{comp.}$	-137.7	-206.7	-338.9	-255.7
R+H$_{PCA}$	24.8	8.4	-58.4	22.4

Table 6: Average relative improvement in % w.r.t RASTA-PLP with DEQ for the different feature types when the specified noise types at SNR values ranging from $-5\,\mathrm{dB}$ ... inf were added and the DEQ was applied. The cumulative distributions were estimated only from the current utterance.

performance. This is in particular the case when using the HIST features where DEQ was applied after the preprocessing or at the final stage. In both cases, when using many utterances (120) to estimate the statistics or only the current utterance, the combination of RASTA-PLP$_{DEQ}$ and HIST$_{preproc.}$ outperforms RASTA-PLP$_{DEQ}$. Yet only when estimating the statistics from many utterances the performance of the combination of RASTA-PLP$_{DEQ}$ and HIST$_{preproc.}$ also clearly outperforms the combination of both features without any equalization. When using only the current utterance to estimate the statistics this is only the case for low SNR levels ($<$ 10 dB). From this we conclude, that one utterance provides not enough information to reliably estimate the statistics. Further research is necessary to understand why the equalization at intermediate processing steps has such a strong negative effect on HIST features.

6. References

[1] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1 – 15, 1997.

[2] H. Hermansky and N. Morgan., "Rasta processing of speech," in *IEEE Trans. on Speech and Audio Processing*, vol. 2, October 1994, pp. 587–589.

[3] A. Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," in *IEEE Trans. on Speech and Audio Processing.*, vol. 13, no. 3, May 2005, pp. 355–366.

[4] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Recognition of noisy speech: a comparative survey of robust architecture and feature enhancement." in *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, February 2009, pp. 1–17.

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

[6] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," in *J. Acoust. Soc. Am.*, vol. 5, no. 131, May 2012, pp. 4134–51.

[7] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-temporal analysis of speech using 2-d gabor filters," Proc. Interspeech, 2007.

[8] S. Ravuri and N. Morgan, "Using Spectro-Temporal Features to Improve AFE Feature Extraction for ASR," in *Proc. Interspeech*, 2010.

[9] G. Sivaram, S. Nemala, N. Mesgarani, and H. Hermansky, "Data-driven and feedback based spectro-temporal features for speech recognition," *Signal processing Letters, IEEE*, vol. 17, no. 8, pp. 957 – 960, Nov. 2010.

[10] H. Wersing and E. Körner, "Learning optimized features for hierarchical models of invariant recognition," *Neural Computation*, vol. 15, no. 7, pp. 1559–1588, 2003.

[11] X. Domont, M. Heckmann, F. Joublin, and C. Goerick, "Hierarchical spectro-temporal features for robust speech recognition," in *Proc. ICASSP*, 2008.

[12] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Communication*, vol. 53, no. 4, pp. 736–752, 2011.

[13] M. R. Schädler and B. Kollmeier, "Normalized spectro-temporal gabor filter bank features," in *Proc. Interspeech*, 2012.

[14] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filterbank," in *Tech. Rep. 35, Apple Computer, Inc.*,, 1993.

[15] C. Kim and R. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *In: Proc. ICASSP. IEEE, Dallas, TX.*, 2010, pp. 4574–4577.

[16] P. Comon, "Independent component analysis: A new concept?" *Signal Process.*, vol. 36, pp. 287–314, 1994.

[17] P. O. Hoyer, "Non-negative matrix factorization with sparness constraints," *Journal of machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[18] R. G. Leonard, "A database for speaker independent digit recognition," *In Proc. ICASSP*, vol. 9, 1984.

[19] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems,," *Speech Communication*, vol. 12, no. 3, pp. 247–252, 1993.

[20] S. Young and al., "The htk book," *Cambridge*, 2006.

[21] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, Paris(France), September 2000.