Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Robust spectro-temporal speech features with model-based distribution equalization

Samuel Kevin Ngouoko Mboungueng, Martin Ernst Heckmann, Britta Wrede

2013

Preprint:

This is an accepted article published in Proc. Int. Workshop on Image and Audio Analysis for Multimedia Interactive Services. The final authenticated version is available online at: https://doi.org/[DOI not available]

Robust Spectro-Temporal Speech Features with Model-Based 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

Previously, we applied a distribution equalization on our HIerarchical Spectro-Temporal (HIST) features using distributions estimated from histogram of one or several utterances. Although a performance increase could be observed in both cases, we noticed low performance improvement when estimating the distribution only from one utterance. The aim here is to determine a parametric distribution from few data samples which gives the highest probability of producing the observed data considering different models. Afterwards, we perform a distribution equalization based on the estimated model after each feature extraction step of our HIST feature extraction framework. We compare the performance of the HIST features with those of conventional spectral features (RASTA-PLP), when a corresponding distribution equalization has been carried out using the TIDigits database and different noise types. By combining the HIST and RASTA-PLP features we achieved on the one hand better performance than spectral features (RASTA-PLP), on the other hand better performance than the same feature combination with distribution equalization obtained directly from the histogram without imposing a parametric distribution model.

1. Introduction

Speech recognition scores of machine decrease significantly in comparison to humans in difficult environments [1], e.g. when the noise exhibits nonstationary characteristics. Thus, common spectral speech features as the Mel Frequency Cepstral Coefficients (MFCCs) or RelAtive SpectrAl (RASTA) features [2] show good performance in clean conditions but this performance strongly deteriorates in the presence of noise. However, *Spectro-temporal features* provided promising results in severe environments [3, 4]. 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 are extracted using Gabor filters [3, 5, 6]. In order to enhance the feature representation and consequently improve their performance in difficult environments several normalization methods in the feature space as *Cepstral Mean Subtraction* (CMS) or *Mean Variance Normalization* (MVN) have been proposed [7, 8]. These methods improve the performance of MFCCs [7] or Gabor-based spectro-temporal [8] features in difficult environments. 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 [7, 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 citeHistDomont, Heckmann2010. 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.

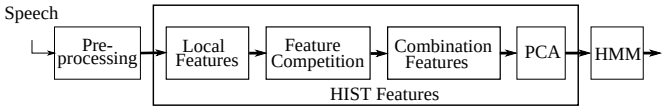


Figure 1: Overview of the feature extraction process [4].

We applied a distribution equalization to our HIST features using cumulative distribution estimated from histogram of one utterance. Only a low performance improvement could be observed compared to results obtained with features without any equalization [9]. Instead of estimating the distribution from the histogram, we will determine a parametric distribution model from the few utterance samples whose parameter estimates provide the highest probability of producing the observed data.

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 model-based distribution equalization. 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, 4]:

1- In the first step the input speech signals are transformed into the spectro-temporal domain using a *Gammatone filterbank* [12], 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 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 [**?**].

2- 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) [13]. 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.

3- In the second layer complex or combination features are obtained by combining local features over a large frequency range. The features are learned using Non-Negative Sparse Coding[14].

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

3. Model-based Distribution Equalization

Let x_0 be a variable having 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 [7]:

$$p_1(x_1) = p_0(G(x_1))\frac{\partial G(x_1)}{\partial x_1}, \tag{1}$$

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

$$\begin{aligned} C_0(x_0) &= \int_{-\infty}^{x_0} p_0(x_0')dx_0' \\ &= \int_{-\infty}^{F(x_0)} p_0(G(x_1'))\frac{\partial G(x_1')}{\partial x_1'}dx_1' \\ &= \int_{-\infty}^{F(x_0)} p_1(x_1')dx_1' \\ &= C_1(F(x_0)) \end{aligned} \tag{2}$$

and therefore, the desired transformation $x_1 = F(x_0)$ is obtained from (2) as

$$x_1 = F(x_0) = C_1^{-1}[C_0(x_0)] = C_{\text{ref}}^{-1}[C_0(x_0)], \tag{3}$$

where $C_0(x_0)$ is the current cumulative distribution and C_{ref}^{-1} is the inverse of the reference distribution function. Instead of determining the CDFs using the histogram as in [9], we estimate the parametric distribution model of each utterance which fits the corresponding empirical CDF. Among several methods for determining the empirical CDF we chose the Hazen method [15]:

$$cdf(x_0) = \frac{i - 0.5}{N}, \text{ with } i = 1, 2, ..., N \tag{4}$$

where N is the number of samples of the vector x_0. Furthermore, we estimate the distribution parameter from the utterance samples by maximizing the *Log Likelihood function* $l(a, b)$:

$$\max_{a,b}\ l(a,b) = \max_{a,b} \sum_{i=1}^{N} \log[f_{X_0}(x_0; a, b)] \tag{5}$$

where $f_{X_0}(x_0; a, b)$ is the distribution model whose parameters a and b have to be estimated such that the observed data can be reproduced with a high probability. We considered three different distributions (normal, laplace and lognormal) whose parameters have been estimated during the aforesaid process.

The distribution equalization is characterized by the fact that nonlinear distortions are compensated, although its efficiency strongly depends on the quality of the estimated parametric model. 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 [16], a database containing 326 speakers each pronouncing 77 digits sequences. We corrupted the data with different types of noise (white, factory, babble and car) from the Noisex database [17] at Signal-to-Noise Ratio (SNR) levels from -5dB . . . inf (clean signal). The recognition is performed with Hidden Markov Models (HMMs) trained on clean signals with HTK [18] using, as defined in the Aurora-2 experimental framework [19], whole 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 model-based distribution equalization procedure is applied during both training and recognition. We use the 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

	white	factory	babble	car
RASTA-PLP$_{mDEQ_\text{lapl}}$	-3.8	-2.7	3.1	-3.6
R+H$_{preproc_\text{norm}}$	20.8	16.6	-76.8	29.9
R+H$_{preproc_\text{lapl}}$	0.2	-8.6	-20.9	18.1
R+H$_{preproc_\text{logn}}$	19.7	14.7	-36.2	32.9
R+H$_{loc_\text{norm}}$	8.3	-20.5	-142.0	14.5
R+H$_{loc_\text{lapl}}$	13.1	9.2	-10.5	21.5
R+H$_{loc_\text{logn}}$	-16.8	-27.7	-33.5	-41.3
R+H$_{comp_\text{norm}}$	-22.8	-91.4	-335.5	-36.8
R+H$_{comp_\text{lapl}}$	-64.9	-123.2	-362.6	-74.9
R+H$_{comp_\text{logn}}$	-42.3	-124.9	-312.9	-32.2
R+H$_{PCA_\text{norm}}$	14.1	-13.5	-174.1	-1.3
R+H$_{PCA_\text{lapl}}$	15.1	1.8	-77.1	11.7

Table 1: Average relative improvement in % w.r.t RASTA-PLP with model-based DEQ using normal model. The distribution models (normal, laplace and lognormal) have been applied after each processing step of HIST feature extraction and after RASTA-PLP features.

	white	factory	babble	car
RASTA-PLP$_{baseline}$	43.1	41.2	35.1	19.7
RASTA-PLP$_{DEQ}$	26.3	26.8	29.4	12.5
RASTA-PLP$_{DEQ_m}$	26.3	26.7	28.6	11.4
R+H$_{baseline}$	24.4	27.1	71.9	15.4
R+H$_{preproc}$	22.9	23.6	31.3	7.5
R+H$_{preproc_m}$	21.4	22.7	39.9	7.2
R+H$_{loc}$	26.7	34.3	48.6	14.5
R+H$_{loc_m}$	22.7	24.2	30.0	8.3
R+H$_{comp.}$	60.6	61.2	79.7	44.9
R+H$_{comp._m}$	35.9	44.1	66.2	19.3
R+H$_{PCA}$	22.6	28.2	43.1	11.4
R+H$_{PCA_m}$	22.6	26.8	40.0	10.2

Table 2: Average word error rates in % for the different feature types when the specified noise types were added over the SNR range from $-5\,\text{dB}\ldots$ inf. The DEQ from histogram and model-based DEQ were applied.

since they exhibit superior performance compared to MFCC features [4]. We used 120 utterances for determining references distributions or model parameters. In the tables and figures below, the abbreviation R+H is referred to as RASTA-PLP + HIST; and suffixes $_{preproc}$, $_{loc.}$, $_{comp.}$ and $_{PCA}$ are referred to as distribution equalization (estimated from histogram) after preprocessing, local, complex or PCA features, respectively. The same suffixes followed by the subindex $_m$ means that the model-based DEQ has been applied instead.

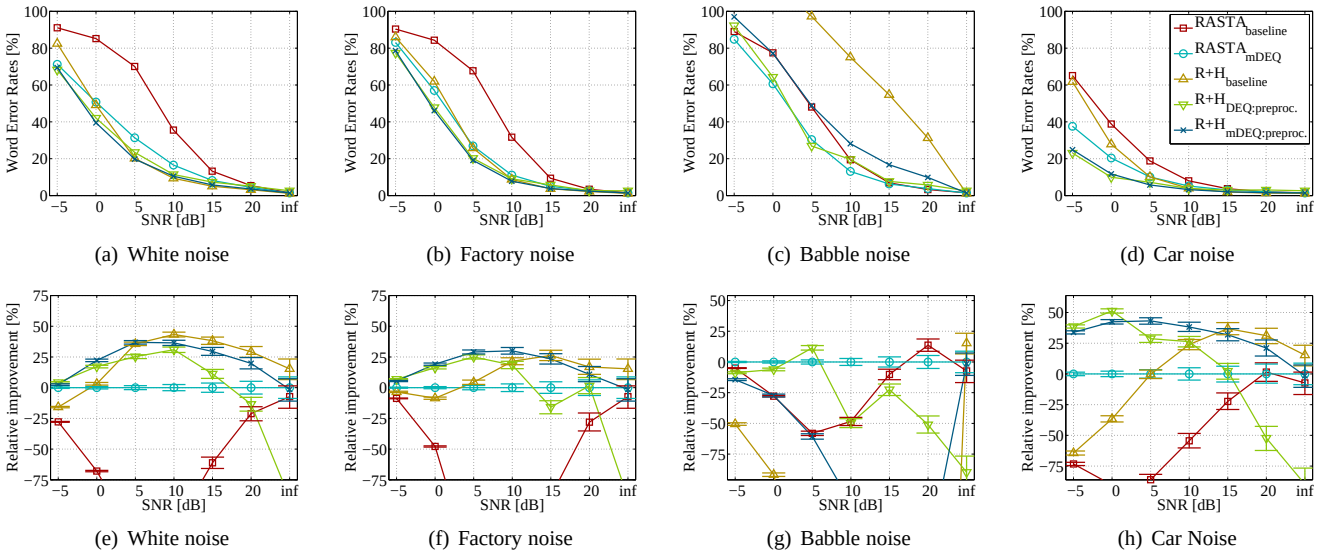
Table 1 shows the average relative performance improvement for each noise type after applying the model-based DEQ considering the different models (normal, laplace and lognormal). Hereby, the average relative improvement is calculated as the average of the relative improvements for each SNR level from $-5\,\text{dB}\ \ldots$ inf as shown in Fig. 2(e, f, g, h). Comparing the results for each processing step without considering babble noise, we find out that the normal model provides the best results after preprocessing and complex feature extraction steps, and the laplace model after local and PCA features. These results will be considered below for the further evaluations.

Tables 2, 3 show the average recogniton performance and the average relative performance improvement for each type of noise, respectively. Features without equalization are referred to as *baseline*. First, the results demonstrate that the performance of RASTA-PLP with model-based DEQ compared to RASTA-PLP with DEQ is improved for each type of noise. We notice a relative improvement of about 19% on average for all types of noise. Further, the results obtained show on the one hand that applying the model-based DEQ after the preprocessing and PCA features a performance improvement is achieved for all noise types except babble noise due to its non stationary characteristic which may not fit the framework where a parametric model normalization is accomplished; on the other hand that applying it after the local features a performance improvement is observed for all noise types. Without considering the babble noise, we reached a relative improvement of about 39.1%, 33.2% or 28.6% on average for the three noise types applying the model-based DEQ, and an average of about 26.2%, -8.7% or 18.5% applying simply the DEQ after preprocessing, local- or PCA features, respectively.

However, when applying the model-based DEQ in the intermediate stage e.g. after the complex feature extraction, we get indeed an improvement compared to the DEQ, but the recognition scores are still inferior to our benchmark the RASTA-PLP with DEQ. We could also observe that the best results are obtained using the model-based DEQ after the preprocessing. By comparing these best results with the combination of the HIST and RASTA-PLP features without any equalization, we can make twofold remarks (see Fig. 2): firstly, the performances of the combination of the HIST and RASTA-PLP features with model-based DEQ are slightly inferior to those without any equalization for $\text{SNR} \geq 10\,\text{dB}$. Nevertheless, we achieved an improvement in this SNR range using the model-based DEQ compared to the DEQ from histogram. Secondly, the performances of the combination of the HIST and RASTA-PLP features with model-based DEQ and DEQ from histogram are comparable for $\text{SNR} < 10\,\text{dB}$, and they are superior to the performances of the combination of HIST and RASTA-PLP features without any equalization.

5. Conclusion

In this contribution we investigated a model-based distribution equalization on HIST and RASTA-PLP features using each utterance. Different models have been considered and we observed that the feature performances depend also on model types. Estimated parametric model distribution seems to better model observed data with few samples than model obtained from histogram. The results show that the parametric model distribution outperforms distribution from histogram especially when the $\text{SNR} \geq 10\,\text{dB}$ whereas the



(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 model-based DEQ (e, f, g, h) of the features when the statistics were estimated only from the current utterance.

	white	factory	babble	car
RASTA-PLP$_{baseline}$	-40.6	-48.0	-3.7	-15.5
RASTA-PLP$_{DEQ_m}$	15.8	16.8	16.1	27.0
R+H$_{baseline}$	32.3	23.6	-253.9	19.8
R+H$_{preproc}$	21.2	19.2	-3.0	38.1
R+H$_{preproc_m}$	34.5	31.6	-47.7	51.3
R+H$_{loc}$	3.0	-26.8	-86.9	-2.5
R+H$_{loc_m}$	28.7	25.9	7.9	45.0
R+H$_{comp.}$	-137.7	-206.7	-338.9	-255.7
R+H$_{comp._m}$	-7.6	-61.5	-253.3	-9.4
R+H$_{PCA}$	24.8	8.4	-58.4	22.4
R+H$_{PCA_m}$	30.6	19.2	-47.3	36.0

Table 3: Average relative improvement in % w.r.t RASTA-PLP with DEQ for the different feature types when the specified noise types were added over the SNR range from $-5\,\mathrm{dB}\ldots$ inf. The statistics were estimated only from the current utterance.

performances are close to each other for $\mathrm{SNR} < 10\,\mathrm{dB}$. In particular for medium to high SNR levels a unimodal distribution equalization might not be optimal. We think it is interesting to explore the multimodal distribution equalization where the models of speech and noise can be separated.

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

[4] 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.

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

[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] 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.

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

[9] S. K. Ngouoko M., M. Heckmann, and B. Wrede, "Spectro-temporal features with distribution equalization," in *SAPA Workshop*, 2012.

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

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

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

[15] R. J. Hyndman and Y. Fan, "Sample quantiles in statistical packages," *American Statistician*, vol. 50, pp. 361 – 365, 1996.

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

[17] 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.

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

[19] 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.