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

Model based distribution equalization applied on spectro-temporal speech features

Samuel Kevin Ngouoko Mboungueng, Martin Ernst Heckmann, Britta Wrede

2013

Preprint:

This is an accepted article published in Proc. Jahrestagung Deutsche Gesellschaft für Akustik (DAGA). The final authenticated version is available online at: https://doi.org/[DOI not available]

Model based distribution equalization applied on spectro-temporal speech features

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.*

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, standard 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. However, *spectro-temporal features* achieved promising results in such situations [3, 4]. Unlike standard features, they are able to detect for instance steady formant transitions in the spectro-temporal representation. Most of them use Gabor filters [5], whereas we developed features inspired by a hierarchical system for visual object recognition [6] . We refer to them as Hierarchical Spectro-Temporal (HIST) features with their extraction scheme depicted in Fig. 1 [4].

Previously, we applied a distribution equalization on our Hierarchical Spectro-Temporal (HIST) features using empirical distributions estimated from one or a reduced set of utterances [7]. 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 whose parameter estimates give the highest probability of producing the observed data. Afterwards, we perform a distribution equalization based on the estimated model after some feature extraction steps of our HIST feature extraction framework.

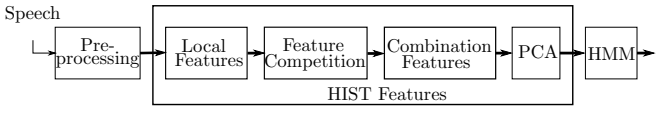


Figure 1: Overview of the feature extraction process [4].

Model-based Distribution Equalization

A detailed description of the HIST feature extraction can be found in [4]. Let x_0 be a variable having a probability density function (pdf) $p_0(x_0)$. A distribution equalization (DEQ) is defined as a transformation $x_1 = F(x_0)$ that converts the pdf $p_0(x_0)$ into the reference pdf $p_1(x_1) = p_{\text{ref}}(x_1)$. Since the cumulative distribution function (CDF) is a monotonically increasing function whereas the inverse can be determined easily, the desired transformation is obtained according to the expression [7]:

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

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 [7], 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 [8]:

$$cdf(x) = \frac{i - 0.5}{N}, \text{ with } i = 1, 2, ..., N \qquad (2)$$

where N is the number of samples of the vector x. Furthermore, we estimate the distribution parameter from the utterance samples by maximizing the *Likelihood function* such that the observed data can be reproduced with a high probability. We considered a normal distribution whose parameters have been estimated during the aforesaid process.

Results

For the evaluation we use TIDigits [9], 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 [10] at Signal-to-Noise Ratio (SNR) levels from $-5\text{dB} \ldots \text{inf}$ (clean signal). The recognition is performed with Hidden Markov Models (HMMs) trained on clean signals with HTK using, as defined in the Aurora-2 experimental framework [11], 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 since they exhibit superior performance compared to MFCC features [4]. In the table and figures below, the abbreviation R+H is referred to as RASTA-PLP + HIST; and suffixes $_{preproc}$, $_{comp.}$ and $_{PCA}$ are referred to as distribution equalization (estimated from histogram) after preprocessing, complexe features or PCA, respectively. The same suffixes followed by the subindex $_m$ means that the model-based DEQ has been applied instead. As mentioned before we chose the normal distribution model whose parameters have to be estimated for model-based DEQ.

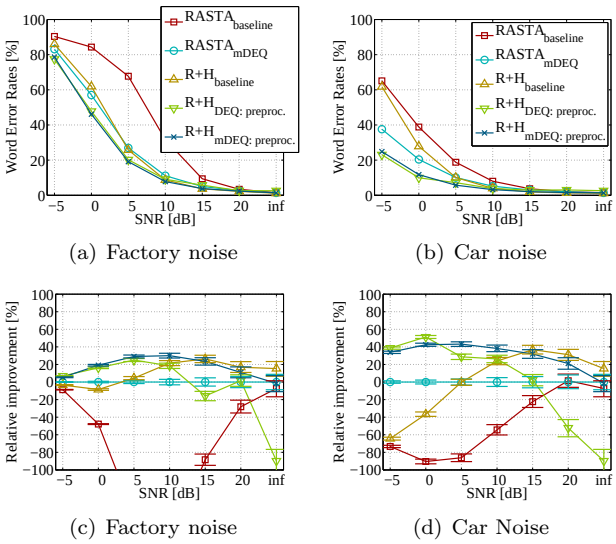
Table 1 shows the average relative performance improvement for each type of noise. Hereby, the average relative improvement is calculated as the average of the relative

improvements for each SNR level from −5 dB ... inf as shown in Fig. 2(c, d). 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 that applying the model-based DEQ after the preprocessing and PCA a performance improvement for all noise types except babbel noise is achieved by combining the HIST and RASTA-PLP features. Without considering the babble noise, we reached a relative improvement of about 39.1% or 17.4% on average for the three noise types applying the model-based DEQ, and an average of about 26.2% or 18.5% applying simply the DEQ after preprocessing or PCA, respectively. The achieved improvement after the PCA with model-based DEQ is still inferior to the DEQ from histogram. However, when applying the model-based DEQ in the intermediate stage e.g. after the complexe 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. Considering these best results and comparing them to the combination of the HIST and RASTA-PLP features without any equalization, we can make twofold remarks (see Fig. 2): on the one hand for SNR $\geq$ 10 dB the performances of the combination of the HIST and RASTA-PLP features with and without model-based DEQ are comparable. However, we achieved an improvement in this SNR range using the model-based DEQ compared to the DEQ from histogram. On the other hand, for SNR $<$ 10 dB the performances of the combination of the HIST and RASTA-PLP features with model-based DEQ and DEQ are comparable, but the performances of the combination of HIST and RASTA-PLP features without any equalization are inferior.

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 SNR $\geq$ 10 dB whereas the performances are closed to each other for SNR $<$ 10 dB. Further experiments will be carried out using other types of distribution models in order to evaluate whether a better data representation will lead to better recogniton scores.

	white	factory	babble	car
RASTA-PLP$_{baseline}$	-40.6	-48.0	-3.7	-15.5
R+H$_{baseline}$	32.3	23.6	-253.9	19.8
RASTA-PLP$_{DEQ_m}$	15.8	16.8	16.1	27.0
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$_{comp.}$	-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$_{PCA_m}$	27.1	4.2	-127.2	20.8

Table 1: 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 dB ... inf were added and the DEQ from histogram and model-based DEQ were applied.



(a) Factory noise (b) Car noise

(c) Factory noise (d) Car Noise

Figure 2: Word error rates (a, b) and relative improvement w.r.t RASTA-PLP with model-based DEQ (c, d) of the features when the cumulative distribution or model was estimated only from the current utterance.

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

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

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

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

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

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

[10] 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.

[11] 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.