Generalization performance of spetro-temporal speech features

Martin Heckmann
Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany

Introduction

Despite the fact that the dynamic aspects of speech are very important, conventional speech features as Mel Ceptstral Coefficients (MFCCs) [1] and RelAtive SpecTrAl Perceptual Linear Predictive (RASTA-PLP) features [2] capture only stationary spectral information. We could previously show that a combination of conventional speech features with spectro-temporal speech features yields to improved recognition results in noisy speech [3, 4]. We termed those latter features as Hierarchical Spectro-Temporal (HIST) features. They consist of two layers, the first capturing local spectro-temporal variations and the second integrating them into larger receptive fields (compare Fig. 1). This layout was inspired by a recently proposed system for visual object recognition [5]. On the first layer we apply ICA (Independent Component Analysis) and in the second layer we apply different learning algorithms, detailed below. Finally we use a Principal Component Analysis (PCA) to orthogonalize the features and further reduce their dimensionality followed by a Hidden Markov Model (HMM) for the recognition.

In addition to investigating three different learning algorithms applied on the second layer of the hierarchy in this paper we also investigate the impact of the database used for learning on the performance. We evaluate the generalization capabilities of the features via deploying different databases during learning of the features and recognition tests.

Combination Feature Learning

We investigate different approaches to learn the receptive fields $\boldsymbol{w}_k^{(2)}$ on the second layer of our feature hierarchy.

Non-negative Matrix Factorization

In Non-negative matrix factorization (NMF) the input data to be represented, the basis functions of the factorization, and the weights at which the basis functions are applied are all positive. For the learning we cut out patches $\boldsymbol{P}$ of length $\Delta = 40\,\text{ms}$ of the first layer activations $\boldsymbol{c}_l^{(1)}$. From these patches we learn $n_2 = 50$ combination features by minimizing the cost function given in (1) [6], where $\boldsymbol{P}_i$ is a tensor representing the n_1 layers of the i-th patch, the $\boldsymbol{w}_k^{(2)}$ are n_2 non-negative tensors each of them containing the n_1 receptive fields $\boldsymbol{w}_{l,k}^{(2)}$, and the $\alpha_{k,i}$ are nonnegative reconstruction factors.

Figure 1: Overview of the feature extraction framework.

Non-negative Sparse Coding

Non-negative Sparse Coding (NNSC) is an extension of NMF which puts a constraint on the coefficients to obtain an efficient use of the basis. This is obtained via a so called sparsity term λ which favors reconstructions of $\boldsymbol{P}$ with a sparse usage of the basis $\boldsymbol{w}^{(2)}$ via a minimization of the weights (see (1))

Weight Coding

The two learning algorithms presented so far where completely unsupervised, i. e. they are not using any class specific information. However, basic functions which mainly capture the information characteristic for a specific class could be beneficial. By introducing a term κ in the cost function (1) which penalizes correlations between projections of patches $\boldsymbol{P}_i$ and $\boldsymbol{P}_j$ from two different classes with the same basis function $\boldsymbol{w}_2^k$ the learning is class specific, and hence not unsupervised anymore [7] (see (1)). $q(i)$ denotes the class label of $\boldsymbol{P}_i$, $n_{q(i)}$ is the number of samples in the class of $\boldsymbol{P}_i$, and T denotes the transpose operator.

Results

The features were trained on the TIMIT [8] corpus, containing phonetically rich sentences. As benchmark we also extracted RASTA-PLP features [2]. We performed recognition experiments on a noisy continuous digit recognition task where we added to TIDigits [9] white noise, noise recorded in a factory and in a car and babble noise at Signal to Noise Ratios (SNRs) ranging from $-5\,\text{dB}\ldots$ inf, i. e. also keeping the clean signal. The HMMs were trained with HTK [10] using whole word HMMs containing 16 states without skip transitions and a mixture of 3 Gaussians with a diagonal covariance matrix per state.

The results in Table 1 show that the combination of HIST and RASTA-PLP features improves results, especially for medium and high SNR values. To better asses this we also calculated the relative improvements of the feature combination compared to RASTA-PLP features alone (compare Fig. 2). This reveals that the combination of HIST and RASTA-PLP features independent of the learning algorithm improves results for all noise types and SNR levels with the exception of babble noise. We have seen this unfavorable behavior of the HIST features in babble noise already previously [4]. Via additional experiments we concluded that the reason for this is the very high sensitivity of the HIST features to speech, also mixtures of different speech signals as in babble noise. A remedy to this is the insertion of babble noise also in the training phase [4].

$$E = \underbrace{\underbrace{\sum_i \|\boldsymbol{P}_i - \sum_{k=1}^{n_2} \alpha_{k,i} \boldsymbol{w}_k^{(2)}\|^2}_{NMF} + \lambda \sum_i \sum_{k=1}^{n_2} |\alpha_{k,i}|}_{NNSC} + \underbrace{\frac{1}{2}\kappa \sum_k \sum_{\substack{i,j \\ q(i)\neq q(j)}} \frac{\boldsymbol{w}_k^{(2)^T}\boldsymbol{P}_i}{n_{q(i)}} \frac{\boldsymbol{w}_k^{(2)^T}\boldsymbol{P}_j}{n_{q(j)}}}_{\text{Weight Coding}} \tag{1}$$

	white	factory	babble	car
RASTA-PLP	43.1	41.0	35.0	19.5
HIST-NMF	41.2	39.4	55.7	16.3
HIST-NNSC	44.4	42.6	64.7	16.3
HIST-WC	40.2	38.6	58.4	16.4
RASTA-PLP+HIST-NMF	32.9	32.5	49.5	11.4
RASTA-PLP+HIST-NNSC	38.0	38.4	59.7	12.8
RASTA-PLP+HIST-WC	35.9	35.0	58.4	12.4
RASTA-PLP+HIST-NMF$_{\text{TI}}$	27.9	30.3	49.0	10.6
RASTA-PLP+HIST-NNSC$_{\text{TI}}$	30.1	31.4	44.8	11.6

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

In this first experiment NMF shows the best performance. NNSC and WC perform very similar to NMF for medium to high SNR values but show clear inferior behavior at low SNR values. Thereby the performance of WC lies in between those of NMF and NNSC.

In a second experiment we investigated to what extent the database used in the learning of the features influences the performance. Therefore, we also applied the TIDigits database for the learning of the features. In contrast to the previous experiment now the database used for learning the features and for evaluating their performance match. Results of this experiment are given in Table 1 as well as Fig. 3 and indicated by the subscript TI. As one can see in case of the NMF results obtained when learning the features on TIDigits are very similar to those obtained when learning them on TIMIT. Overall the results in the case where the database used for learning the features and performing the recognition experiments match, i. e. in both cases TIDigits, are slightly better. However, the performance of NNSC improved significantly in this matched learning condition. In this case the difference between NMF and NNSC is only small. As the two databases we compared during learning of the features cover a quite different domain, we conclude that the information captured by the HIST features when using NMF for learning is indeed not database but speech specific.

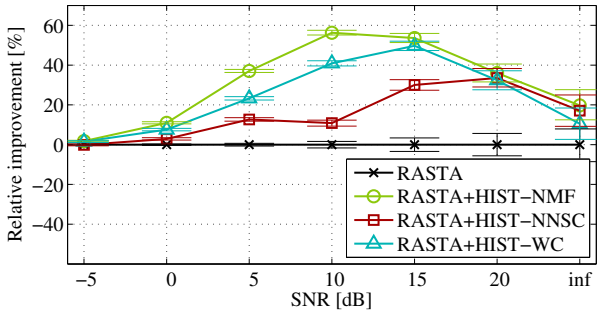

Figure 2: Relative improvements compared to RASTA-PLP features when factory noise was added to the test set. The bars indicate the 95% confidence intervals calculated according to [11].

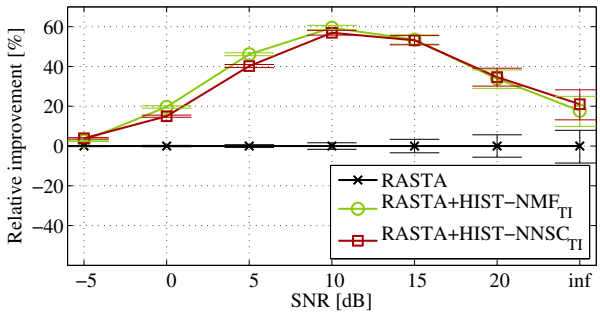

Figure 3: Relative improvements of the features trained on TIDigits compared to RASTA-PLP features when factory noise was added to the test set. The bars indicate the 95% confidence intervals calculated according to [11].

References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 28, no. 4, pp. 357–366, 1980.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans Speech and Audio Proc.*, vol. 2, no. 4, pp. 578–589, 1994.

[3] X. Domont, M. Heckmann, F. Joublin, and C. Goerick, "Hierarchical sectro-temporal features for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, Las Vegas, NV, 2008, pp. 4417–4420.

[4] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Communication*, 2011.

[5] H. Wersing and E. Körner, "Learning Optimized Features for Hierarchical Models of Invariant Object Recognition," *Neural Computation*, vol. 15, no. 7, pp. 1559–1588, 2003.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[7] S. Hasler, H. Wersing, and E. Körner, "Combining reconstruction and discrimination with class-specific sparse coding," *Neural Computation*, vol. 19, no. 7, pp. 1897–1918, 2007.

[8] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren, *DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM*, Philadelphia, 1993.

[9] R. Leonard, T.I. Incorporated, and T. Dallas, "A database for speaker-independent digit recognition," in *Int. Conf. Acoustics, Speech, and Signal Proc. (ICASSP)*, San Diego, CA, 1984, vol. 9, IEEE.

[10] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University, 1995.

[11] J.M. Vilar, "Efficient computation of confidence intervals for word error rates," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, Las Vegas, NV, 2008, pp. 5101–5104, IEEE.