Investigating the Complementarity of Spectral and Spectro-temporal Features

Martin Heckmann[1], Xavier Domont[1,2], Frank Joublin[1], Christian Goerick[1]

[1] *Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Email: {firstname.lastname}@honda-ri.de*

[2] *Technische Universität Darmstadt, Regelungsth. u. Robotik, D-64283 Darmstadt, Email: xavier.domont@rtr.tu-darmstadt.de*

Introduction

Most common speech features as Mel Ceptstral Coefficients (MFCCs) and RelAtive SpecTrAl Perceptual Linear Predictive RASTA-PLP features use only spectral information. However, from measurements in the mammalian auditory cortex it is known that the mammalian brain jointly uses spectral and temporal information. To model this we previously developed Hierarchical Spectro-Temporal (HIST) features [1, 2]. They consist of two layers, the first capturing local spectro-temporal variations and the second integrating them into larger receptive fields. This layout was inspired by a recently proposed system for visual object recognition [3].

Potentially spectro-temporal features can better model the relevant speech information as purely spectral features. In this paper we will highlight that the information extracted by the proposed Hist features in fact differs from that of purely spectral information as extracted by Rasta-Plp and Mfcc features and we will show that a combination of both feature types can be used to reduce word error rates in a noisy digit recognition task.

In the following we will first briefly introduce the Hist features. More details on the features and a more thorough analysis can be found in [2]. Next we will perform a covariance analysis between Hist , Rasta-Plp , and Mfcc features. Finally we will show based on recognition results that the complementary information extracted by the Hist features is able to reduce word error rates.

Hierarchical Spectro-Temporal Features

The key elements of our hierarchical feature extraction framework are depicted in Fig. 1. First a preprocessing step performs a transformation in the frequency domain with a Gammatone filterbank and a following enhancement of the formant frequencies via a filtering along the frequency axis [2]. This yields the input to the first layer of our hierarchical feature extraction framework. The corresponding receptive fields, i. e. filter kernels, are learned based on Independent Component Analysis (ICA). The receptive fields in the second layer are learned with Non-negative Sparse Coding (NNSC) [2]. In both cases the learning is completely unsupervised and based on randomly selected patches from the training part of the TIDigits database [4]. The second layer yields 50 coefficients to which we add Delta and Delta-Delta features, yielding a 150 dimensional feature vector. On this feature vector we perform a Principle Component Analysis (PCA) to reduce the dimensionality of the feature space to 39 dimensions and to orthogonalize the features. The resulting features are then fed into a Hidden Markov Model (HMM) for recognition.

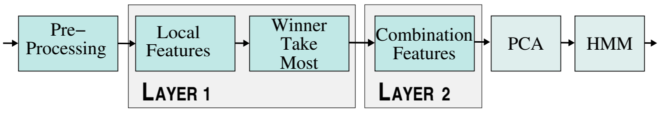

Figure 1: Overview of the feature extraction process.

Experimental Conditions

In the following we want to further investigate the nature of the information extracted by the Hist features. For doing so we perform a covariance analysis to conventional features and perform recognition experiments. But first we want to briefly describe the experimental conditions. We use TIDigits, a database for speaker independent continuous digit recognition. TIDigits contains 326 speakers each pronouncing 77 digit sequences [4]. To this data we added car noise from the Noisex database [5] at varying Signal to Noise Ratios (SNRs). The Hidden Markov Models serving as recognition backend were trained on clean signals with HTK with whole word HMMs containing 16 states without skip transitions and a mixture of 3 Gaussians with a diagonal covariance matrix per state.

Covariance Analysis

As stated above spectro-temporal features should in principle be able to extract information from the speech signal which is not accessible to conventional, purely spectral features. To substantiate this supposition we performed a covariance analysis between the proposed Hist features and conventional Rasta-Plp features [6]. Additionally, we also performed this covariance analysis between the Rasta-Plp features and another commonly used type of spectral features, namely Mfccs [7]. In both cases we calculated the corresponding features from the test set of the TIDigits database without any additional background noise added [2]. We used 39 Hist features following the PCA step as well as 15 Rasta-Plp and 13 Mfcc features without Deltas and Delta-Deltas. The results of this analysis are shown in Fig. 2. When comparing the two covariance matrices one can clearly see on the rectangular off-diagonal parts in Fig. 2(a) that the correlation between Hist and Rasta-Plp features is much smaller than the correlation between Rasta-Plp and Mfcc features. This demonstrates that Rasta-Plp and Mfcc features extract very similar information and at the same time that the information extracted by the

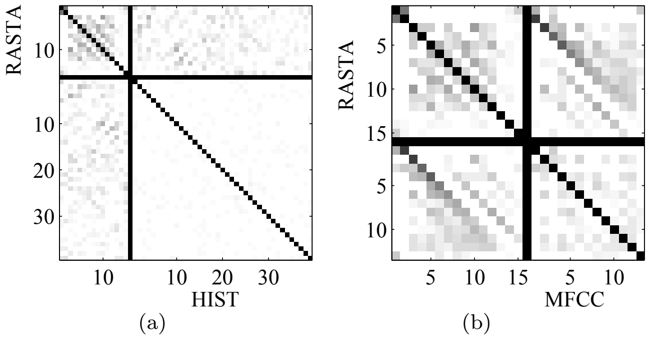

Figure 2: Comparison of covariance matrices between the combination of RASTA-PLP and HIST features (a) and the combination of RASTA-PLP and MFCC features. Prior to the calculation of the covariance matrices we removed the mean and performed a variance normalization of each feature individually. The separation between the two feature sets is highlighted with a black bar.

HIST features is different.

Recognition Scores

The results of the previous covariance analysis showed that the HIST features extract different information than RASTA-PLP and MFCC features but it left open if this information is indeed useful. To investigate this we performed recognition tests on the TIDigits database as described above. From Fig.3 it can first be seen that RASTA-PLP features are more robust against additional background noise than MFCCs. Next it can be observed that for high SNR levels the performance of the HIST features is inferior to RASTA-PLP or MFCC features but at low SNR levels the HIST features show better performance than the other two feature types. Furthermore, the results show that the combination of either HIST with RASTA-PLP or HIST with MFCC features clearly improves the performance. To better evaluate this Fig. 4 shows the relative word error rates with results obtained using RASTA-PLP features as a baseline. The plot shows that combining HIST features and RASTA-PLP or MFCC features yields relative improvements of more than 50% for medium SNR levels. Combining RASTA-PLP and MFCC features on the other hand has only a small positive effect for high SNR levels and is even disadvantageous for low SNR levels.

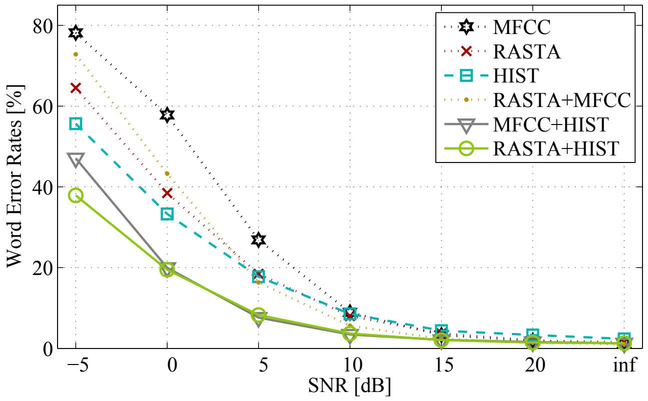

Figure 3: Word error rates (WERs) when car noise was added to the speech data.

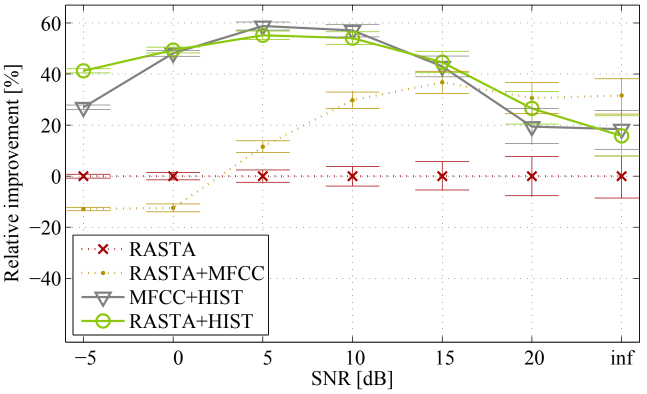

Figure 4: Word error rates (WERs) when car noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervalls calculated according to [8].

Discussion

Based on a covariance analysis between our proposed HIST features and conventional, purely spectral features we could show that our spectro-temporal features capture information which is only weakly correlated with the information extracted by conventional features. Additional speech recognition experiments showed that combining the spectro-temporal features with purely spectral features significantly reduces error rates in noise. From this we conclude that the information captured via the proposed spectro-temporal features is not only different from that extracted by conventional features but also complementary to it.

References

[1] 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, Nevada, 2008, pp. 4417–4420, IEEE.

[2] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, “A hierarchical framework for spectro-temporal feature extraction,” *submitted to Speech Communication.*

[3] 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.

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

[5] 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. 247–251, 1993.

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

[7] 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.

[8] J.M. Vilar, “Efficient computation of confidence intervals forword error rates,” in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*. 2008, pp. 5101–5104, IEEE.