Spectro-Temporal Features with Noise-Adaptive Competition

Samuel K. Ngouoko M.[1], Martin Heckmann[2], Britta Wrede[3]

[1,3] *Research Institute for Cognition and Robotics; Bielefeld University, D-33615 Bielefeld, Germany.*

[2] *Honda Research Institute GmbH, D-63073 Offenbach/Main, Germany.*

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 presence of noise. However, *Spectro-temporal features* gave promising results in such situations [3]. Unlike standard features, they are able to detect for instance steady formant transitions in the spectro-temporal representation. Most of them use Gabor filters [4], whereas we developed features inspired by a hierarchical system for visual object recognition [5] . We refer to them as Hierarchical Spectro-Temporal (HIST) features [6] with their extraction scheme depicted in Fig. 1. Inspired by findings of studies conducted on mammals, which showed the rapid adaptation of the shape of their receptive fields on the current task [7], we introduce a new adaptive feature competition in the HIST feature extraction process. In our approach the features are firstly learned in a noise-free environment, then they are tested by adapting the feature competition to the current noise condition. We propose here not an adaptive preprocessing but an adaptive feature extraction. We show that the adaptation of the feature competition with respect to the noise level improves the recognition rate.

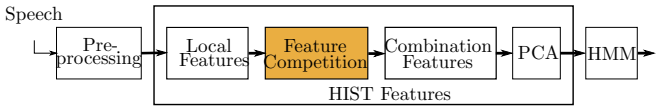


Figure 1: Overview of the feature extraction process [6].

Adaptive Feature Competition

The HIST feature extraction consists of two layers after the preprocessing: In the first layer the local features $q_l^{(1)}(t,f)$ are extracted as the absolute of the 2D convolution of the input spectrogram $\mathbf{S}$ with a set of $l = 1 \ldots n_1$ receptive fields $\boldsymbol{w}_l^{(1)}$. It is followed by a competition between coequal features using the Winner Take Most (WTM) algorithm which cancels the response of less active neurons at the positions (t,f):

$$r_l^{(1)}(t,f) = \begin{cases} 0 & \text{if } \frac{q_l^{(1)}(t,f)}{M(t,f)} < \gamma \\ & \text{or } M(t,f) = 0 \\ \frac{q_l^{(1)}(t,f) - \gamma M(t,f)}{1-\gamma} & \text{else,} \end{cases} \quad (1)$$

where $M(t,f) = \max_l q_l^{(1)}(t,f)$ is the maximal value at position (t,f) over the l neurons and $0 \leq \gamma \leq 1$ is a parameter controlling the strength of the competition [5]. In the second layer features resulting from a combination of local features are extracted, then orthogonalized using the Principal Component Analysis (PCA).

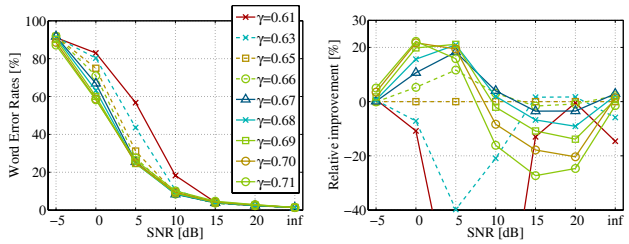
We have previously shown that the competition of local features improves the performance of the HIST features [6]. In these experiments the competition strength was fixed for all acoustical environments. This made the feature extraction independent of the acoustical condition. However, each acoustical environment exhibits its own characteristics. Consequently, we expect improved performance for the feature extraction by adapting the competition strength γ to the acoustical environment. For instance, a strong feature competition is expected when the noise level is high, and vice versa. A straight forward approach to assess the current acoustical environment is to estimate the SNR. Based on this estimate the feature competition can then be set adaptively.

Several methods to estimate the SNR in difficult environments have been proposed. One prominent method is the one proposed by Cohen [8], where the estimated SNR is used for noise spectrum estimation and speech enhancement. We use this method and the corresponding implementation [9] to estimate the SNR.

Results

For the evaluation we use TIDigits [10], a database for speaker independent continuous digit recognition. We corrupted the data with different types of noise (white, factory and car) from the Noisex database [11] at SNR levels from $-5\text{dB} \ldots \text{inf}$ (clean signal). The Hidden Markov Models (HMMs) were trained on clean signals with HTK [12] using 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. We use a combination of HIST and RASTA-PLP features as we obtained previously good recognition scores with such a combination [6]. To evaluate the benefits of the proposed approach, we performed multiple tests.

We found a baseline for our scenario by determining the fixed competition strength γ, which delivers the best performance. For doing so, we varied γ in the range of $0.6 \ldots 0.75$. For a given value of γ we learned the features, trained the HMMs and performed recognition experiments. Thereby $\gamma = 0.65$ gave the best overall performance. Hence, we consider $\gamma = 0.65$ as the benchmark.



(a) Absolut WER (b) Rel. improvement

Figure 2: Word error rates (WER) (a) and relative improvement (b) of the features when the competition strength is varying. Factory noise was added. The features were learned with $\gamma = 0.65$ and are used as baseline in (b).

Noise type \ SNRs		-5	0	5	10	15	20
factory	Opt.	1.7	19.8	21.1	0.0	0.0	0.0
	Adapt.	1.6	19.5	20.5	-2.1	-1.1	0.0

Table 1: Relative improvement in % of the feature performance with respect to the features learned and tested with $\gamma = 0.65$ when factory noise at SNR values ranging from $-5 \ldots 20$ dB we added.

Afterwards, we performed the adaptation of the competition strength dependent on the acoustical environment. More precisely, we used $\gamma = 0.65$ (benchmark, referred to as fixed) for learning the features and training the HMMs and varied γ in small steps during recognition. The results demonstrate that the feature performance strongly depends on the competition strength (Fig. 2). On the one hand, when selecting the optimal γ during recognition for each acoustical environment one notes a significant performance improvement at low SNRs whereas at good SNRs the performance is comparable to that of the fixed γ. The results are shown in Fig. 3 (referred to as optimal).

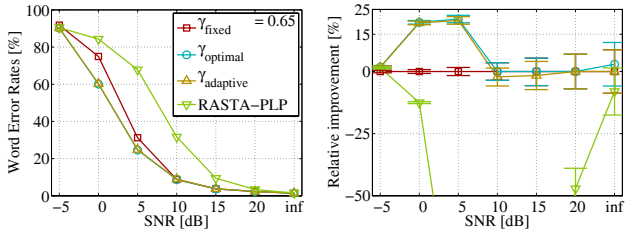
On the other hand, when looking at Fig. 2 one can see that two values for γ are sufficient to obtain almost optimal performance: 0.65 if SNR > 10 dB and a stronger competition ($\gamma = 0.69$) in the other cases. This facilitates the development of an adaptation algorithm. To adaptively set the competition strength, we therefore use the estimated SNR of the input signals, which is the average SNR of all frames containing speech for a given utterance.

With the estimated SNR we control the feature competition using a threshold:

$$\gamma = \begin{cases} 0.65 & \text{if SNR} > 10\,\text{dB} \\ 0.69 & \text{else.} \end{cases} \qquad (2)$$

When using this adaptive process we obtain the results shown in Fig. 3 (referred to as adaptive). As can be seen this adaptive γ improves the feature performance considerably especially at low SNRs compared to the fixed γ setting. The corresponding relative improvements for each SNR level between -5 and 5 dB are given in Table 1. We see very small performance deteriorations due to occasional errors in the SNR estimation and subsequent incorrect setting of the competition factor. Nevertheless, the overall performance of the adaptive process is close to the optimal performance.

By investigating empirical feature competition depending on acoustical environment we determined competition strengths, which could be set dynamically based on an estimate of the SNR. The simulations results revealed that a stronger feature competition at low SNRs provides performance improvement of about 15% compared to fixed competition. Similar results could be obtained using white and car noise.



(a) Absolut WER (b) Rel. improvement

Figure 3: Word error rates (a) and relative improvement (b) of the features using the optimal and adaptive competition strength for each environment. Factory noise was added. The results are compared to the fixed benchmark $\gamma = 0.65$ and RASTA-PLP features alone.

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 Transactions on Speech and Acoustics*, 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 *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

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

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

[6] 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.

[7] J. Fritz, S. Shamma, M. Elhilali, and D. Klein., "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex." *Nature neuroscience*, vol. 6, no. 11, pp. 1213 – 1223, Nov. 2003.

[8] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," in *IEEE Trans. on Speech and Audio Proc.*, 2002.

[9] [Online]. Available: http://webee.technion.ac.il/Sites/People/IsraelCohen

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

[11] 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.

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