# TRAP-based Techniques for Recognition of Noisy Speech

*František Grézl and Jan Černocký*[⋆]

Speech@FIT, Brno University of Technology, Czech Republic,
{grezl,cernocky}@fit.vutbr.cz

**Abstract.** This paper presents a systematic study of performance of TempoRAl Patterns (TRAP) based features and their proposed modifications and combinations for speech recognition in noisy environment. The experimental results are obtained on AURORA2 database with clean training data. We observed large dependency of performance of different TRAP modifications on noise level. Earlier proposed TRAP system modifications help in clean conditions but degrade the system performance in presence of noise. The combination techniques on the other hand can bring large improvement in case of weak noise and degrade only slightly for strong noise cases. The vector concatenation combination technique is improving the system performance up to strong noise.

## 1 Introduction

In recent years, Temporal Pattern (TRAP) based feature extraction has become popular and especially systems combining TRAP with conventional parameters such as MFCC or PLP exhibit good performances [1].

Unlike mostly used features which are based on full spectrum with short time context, temporal pattern (TRAP) features are based on narrow band spectrum with long time context. These features are derived from temporal trajectory of spectral energy in frequency bands in two steps: First, critical band trajectory is turned into band-conditioned class posteriors estimates using nonlinear transformations — neural net. Second, overall class posteriors estimates are obtained by merging all band-conditioned posteriors. The merging is done by another neural net. Overall class posteriors transformed into form required by a standard GMM-HMM decoder are called TRAP (or TRAP-based) features. The fact that the first step of TRAP processing happens in frequency bands should make TRAP-based features robust in frequency selective noise. In [2], TRAP features were tested with the Qualcomm-ICSI-OGI features [3] on the Aurora2 database.

Since, numerous modifications of TRAP features were proposed and tested. The concatenation of several critical bands was tested in [4]. In addition, the

⋆ This work was partly supported by European projects Caretaker (FP6-027231), by Grant Agency of Czech Republic under project No. 102/05/0278 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under projects No. 119/2004, No. 162/2005 and No. 201/2006.

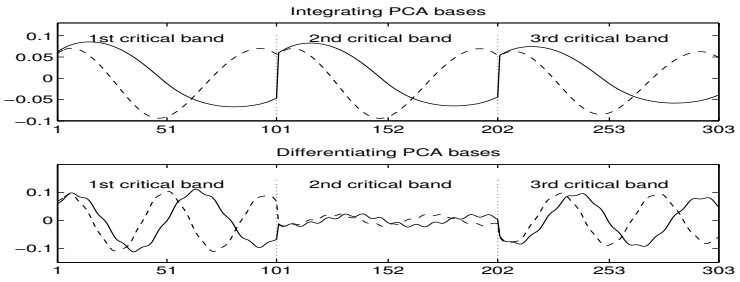


**Fig. 1.** Integrating and differentiating PCA bases

Principal Component Analysis (PCA) was performed on concatenated vectors and the resulting bases were used for dimensionality reduction. It was observed that the PCA bases have similar shapes as the Discrete Cosine Transform (DCT) bases. Further, the majority of the PCA bases for three concatenated critical band energy trajectories have shapes which perform integration and differentiation of individual critical bands (see Fig. 1). In bases performing integration, all parts corresponding to individual bands have similar shape. In bases performing differentiation, the part corresponding to the middle band is close to zero and the shapes for border bands have opposite phases.

In [5] integrating and differentiating of critical bands is applied directly on the critical band spectrogram prior to the temporal pattern selection, creating so called modified temporal pattern (MTRAP). It was shown that one modification (integration or differentiation) itself does not achieve the performance of the basic TRAP system. The combination of two MTRAP systems or MTRAP and basic TRAP system is necessary. Possible combinations are examined in [6] showing the effectiveness of simple vector concatenation technique where temporal patterns from differently modified critical band spectrograms are concatenated on the input of band-conditioned neural net. However, all results are obtained on a small task (digits) on clean telephone speech and the robustness of the proposed improvements to noise was not verified.

We made efforts to evaluate the proposed techniques on noisy speech from Aurora2 database while having only clean training data to see whether these techniques are beneficial also in noisy conditions. The description of experimental setup is given in section 2. The following sections then give the overview of used techniques and results. Section 3 introduces TRAP-based feature extraction, section 4 summarizes the multi-band system and system with critical band spectrogram modification, and section 5 describes the combinations of TRAP systems. Conclusions are given in section 6.

## 2 Experimental setup

The AURORA2 database was designed to evaluate speech recognition algorithms in noisy conditions. The framework was prepared as contribution to the ETSI

STQ-AURORA DSR Working Group [7]. The database consists of connected digits task (11 words) spoken by American English speakers. A selection of 8 different real-world noises has been added to the speech with different signal to noise ratio (SNR). The noises are suburban train, crowd of people, car, exhibition hall, restaurant, street, airport and train station. The noise levels are 20dB, 15dB, 10dB, 5dB, 0dB and -5dB.

The training part of the database consists of 8440 utterances. Only the clean training scenario is used in this work. The test part of the database consists of 4004 sentences divided into 4 sets with 1001 utterances each. One noise with given SNR is added to each subset. There are three test sets: **A** and **B** are noisy conditions containing noises matching (A) and non-matching (B) the noisy training data. The test set **C** is corrupted, in addition to noises, by channel mismatch. Each set thus represents an experiment with unique noisy conditions.

For the training of neural nets, the training part of Aurora2 was forced-aligned using models trained on OGI-Stories database [8]. OGI-Stories were also added to the neural net training set to enrich the phoneme context (in digits, phonemes are occurring in the same context). The target 21 phonemes are those which occur in digits utterances including silence. Other phonemes are not used for training but they create context in TRAP vectors.

The reference recognizer shipped with AURORA was used. The number of results per experiment is given by number of SNR and number of noises. To be able to compare the results from different experiments, we report an average word error rate (WER) for given SNR.

## 3 TRAP-based feature extraction

To obtain the critical band energy trajectory, we have to get the critical-band spectrogram first. This is done by segmentation of the speech into 25 ms frames spaced by 10 ms. Then the power spectrum is computed from each speech frame and integrated by 15 Bark-scaled trapezoidal filters. Finally, logarithm is taken.

In such critical-band spectrogram, the TRAP vector is selected as 101 consecutive frames (center frame $+/-$ 50 frames context) in a given frequency band. The TRAP vector is mean and variance normalized and weighted by Hamming window. In case the PCA or DCT dimensionality reduction is desired, matrix multiplication follows. The resulting vector is then converted into band-conditioned class posteriors by a band-specific **band probability estimator** – a three layer neural net trained to classify the input vector in one of the 21 phonetic classes. All band-conditioned posterior estimates are then concatenated in one vector. Before presenting this vector to the **merger probability estimator** to obtain the overall class posteriors, negative logarithm is taken. Merger probability estimator is also a three layer neural net. Target classes are the same 21 phonemes as for band probability estimator. The block diagram of the TRAP system which converts the critical-band spectrogram to phoneme posteriors estimates is shown in Fig. 2.

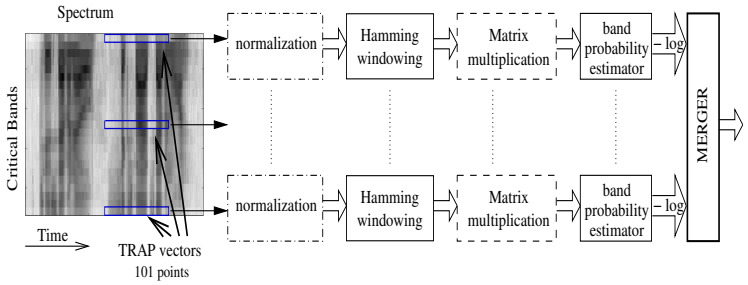


**Fig. 2.** TRAP system for converting the critical band spectrogram to phoneme posteriors. After post-processing, the resulting features are used in standard GMM-HMM recognizer.

The TRAP-based features are obtained from phoneme posteriors obtained by taking the logarithm and PCA decorrelation. This features form an input to the Aurora2 GMM-HMM recognizer.

The resulting features are denoted *basic TRAP* and obtained results are shown in Tab. 1. There is no dimensionality reduction (matrix multiplication) of TRAP vector in basic TRAP features.

## 4 Modifications

### 4.1 Multi-band TRAP system

The multi-band system was proposed in [4]. Three adjacent bands are used as input to the band probability estimator. Frequency shift between two band probability estimators input is one band. Features obtained by this system are denoted as *3b TRAP* and the results are shown on the 3rd line of Tab. 1.

In [4], it was also shown that the dimensionality reduction of concatenated TRAP vectors can further improve the recognition accuracy. We used the neural net training data to compute the PCA bases. The input 303 point long vector was reduced to 150 points. Features obtained by this system are denoted as *3b TRAP + PCA* and the results are shown on the 4th line of Tab. 1.

### 4.2 Critical-band spectrogram modification

According to study presented in [4], it is possible to replace the PCA bases by bases created by concatenating the DCT bases in integrating or differentiating manner (see Fig. 1). In [5], this integration and differentiation was applied directly on the critical band spectrogram using so called *modifying operators*. It was also shown, that replacing the system with integration of critical band spectrogram by the *Basic TRAP system* does not hurt the system performance but rather brings slight improvement. Therefore, we will stick with basic TRAP and differentiation of the critical band spectrogram systems.

The frequency differentiating (FD) operator is a column vector $FD = [1, 0, -1]^T$. The modified critical band spectrogram (MCRBS) is computed as projection of the operator on the original spectrum (CRBS). One point of MCRBS in given time $t$ and in given frequency band $f$ is computed as

$$MCRBS(t, f) = \sum_{i=f-f_c}^{f+f_c} FD(t, i) \times CRBS(i) \quad (1)$$

where $f_c$ is the frequency context of the FD operator (in our case 1).

The processing of the MCRBS is the same as for the Basic TRAP features. Features obtained by this system are denoted as *FD MTRAP* and the results are shown on last line of Tab. 1.

| features | clean | SNR20 | SNR15 | SNR10 | SNR5 | SNR0 | SNR-5 | average |
|---|---|---|---|---|---|---|---|---|
| MFCC | 0.8 | 7.9 | 20.4 | 41.1 | 64.8 | 83.9 | 92.7 | 44.5 |
| basic TRAP | 1.9 | 6.5 | 10.7 | 19.2 | 37.8 | 69.0 | 87.7 | 33.2 |
| 3b TRAP | 1.8 | 5.5 | 9.8 | 20.3 | 42.5 | 74.3 | 89.4 | 34.8 |
| 3b TRAP + PCA | 1.4 | 5.7 | 12.7 | 32.3 | 69.0 | 88.5 | 91.7 | 43.1 |
| FD MTRAP | 2.1 | 7.6 | 15.2 | 33.1 | 63.1 | 84.5 | 90.8 | 42.3 |

**Table 1.** WER [%] for different TRAP-based features.

| combination | clean | SNR20 | SNR15 | SNR10 | SNR5 | SNR0 | SNR-5 | average |
|---|---|---|---|---|---|---|---|---|
| lin ave | 1.5 | 4.4 | 8.9 | 21.7 | 50.7 | 80.5 | 89.8 | 36.8 |
| log ave | 1.3 | 3.8 | 8.2 | 20.6 | 49.4 | 80.4 | 89.6 | 36.2 |
| inv ent th = 1.0 | 1.4 | 3.9 | 7.9 | 18.6 | 44.6 | 78.0 | 89.6 | 34.9 |
| inv ent th = 2.5 | 1.3 | 3.6 | 7.0 | 16.3 | 38.2 | 72.4 | 87.9 | 32.4 |
| vector concat | 1.6 | 4.2 | 7.6 | 15.6 | 36.4 | 70.9 | 88.5 | 32.1 |

**Table 2.** WER [%] of different Basic TRAP and FD MTRAP system combinations.

## 5 System combinations

Combination of TRAP system at different levels is examined in [6]. Simple multi-stream combination and vector concatenation techniques were giving the best results. Here, we apply the concatenation techniques on *Basic TRAP* and *FD MTRAP* systems.

### 5.1 Multi-stream combination

This combination technique combines the final probability estimations from different systems – i.e. outputs of merger probability estimators. The outputs from the TRAP systems are posterior probabilities $P(q_k|\mathbf{x}_t, \theta)$, where the $q_k$ is the $k^{th}$ output class of total $K$ classes, $\mathbf{x}_t$ is the input feature vector at time $t$ and

$\theta$ is set of neural net parameters. The systems have the same targets, thus we can use techniques for posterior probability combination. The resulting posterior probability vector for combining $I$ systems will be $\hat{P}(q_k|\mathbf{X}_t, \mathbf{\Theta})$ where $\mathbf{X}_t$ is the set of all input vectors $\mathbf{X}_t = \{\mathbf{x}_t^1, \mathbf{x}_t^2, \dots \mathbf{x}_t^I\}$ and $\mathbf{\Theta} = \{\theta^1, \theta^2, \dots \theta^I\}$ is the set of all parameters.

First we performed an **average of output probabilities**, which simply averages the outputs belonging to the same class. This combination of *Basic TRAP* and *FD MTRAP* is denoted as *lin ave* and results are shown on the first line in Tab. 2.

Another possibility is to take the **average of logarithm of output probabilities**, which is equivalent to geometric averaging of the linear posteriors. This multi-stream system combination is denoted as *log ave* and results are shown on the second line in Tab. 2.

Finally, we have explored **entropy based combination** inspired by [9], which is actually a weighted average of output probabilities where weights are estimated for each frame individually.

The entropy of $i^{th}$ system outputs at given time $t$:

$$h_t^i = -\sum_{k=1}^{K} P(q_k|\mathbf{x}_t^i, \theta^i) log_2(P(q_k|\mathbf{x}_t^i, \theta^i)) \tag{2}$$

can be used as confidence measure of this system. This information is used for weighting the outputs of different systems. The weight for $i^{th}$ system at time $t$ is

$$w_t^i = \frac{1/h_t^i}{\sum_{i=1}^{I} 1/h_t^i} \tag{3}$$

High entropy means that the posterior probabilities are approaching equal probability for all classes. The stream with high entropy has less discrimination, therefore outputs of such system should be weighted less. The stream with low entropy has higher discrimination and its outputs should be weighted more. This weighting scheme prefers the input stream which has higher disriminability, i.e. is more noise robust.

**Inverse entropy weighting with static threshold** was used in our experiments. If the system entropy at given time is higher than a threshold, the entropy is set to a large value:

$$\tilde{h}_t = \begin{cases} 10000 & : \quad h_t^i > th \\ h_t^i & : \quad h_t^i \leq th \end{cases} \tag{4}$$

If both systems have entropy bigger than threshold $th$, both obtain small (but the same) weight and the output will be equal to the average of both systems. If systems have small entropy $< th$, the output will be given by the weighted average. If only one system has entropy $> th$, this output will be suppressed by the small weight and the output will be given by the system with entropy $< th$. We have tuned the threshold value and best results were obtained with $th = 3.5$.

This combination of *Basic TRAP* and *FD MTRAP* is denoted as *inv ent th = val* where *val* is the threshold value. The results are shown in Tab. 2.

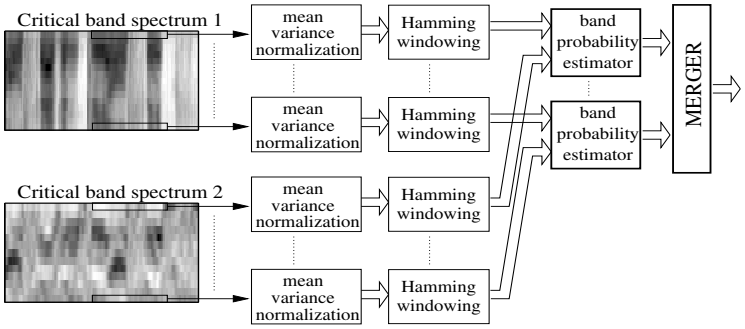


**Fig. 3.** Block diagram of system with vector concatenation

### 5.2 Vector concatenation

The simple way of combination different feature vectors is to directly concatenate them. The concatenation of the TRAP vectors obtained from different critical band spectrograms is done on the input of band probability estimator. It means that all processing (normalization, windowing, DCT) is done for each vector independently. Fig. 3 shows the processing for system with vector concatenation. The results for vector concatenation system combination are denoted *vector concat* and are shown on last line of Tab. 2.

## 6 Conclusions and discussions

The results for standard MFCC features are given for comparison on first line in Tab. 1. MFCC features gain better performance in clean conditions but are more vulnerable in noisy conditions compared to basic TRAP features. Basic TRAP are set as a baseline we compare the other TRAP-based techniques to.

The multi-band TRAP system achieves improvement for clean speech and speech with SNR $>$ 10dB. For stronger noises the performance is inferior to basic TRAP features. We explain this behavior by the fact that concatenation of the TRAP vectors form adjacent critical bands spreads the noise from one band to three band probability estimators. Hence, instead of one impaired band-conditioned posterior estimates in case of basic TRAP system, there are three impaired estimates and the overall estimates suffer.

The PCA dimensionality reduction improves the performance on clean speech in agreement with [4], but the deterioration in noisy cases is severe. This is due to the fact that while doing the matrix multiplication, the change of one point in input vector affects all points of output vector. Thus the noise affects the band estimates much more.

The FD MTRAP features has also very poor performance compared to the basic TRAP but we expect them to help in combination.

The *lin ave* and *log ave* multi-stream combination are able to achieve better performance for weak noises with SNR $<$ 10dB, but the results for stronger noises

are badly affected by the system which is more vulnerable to the noise. This is – to some extent – solved by the inverse entropy based combination. By increasing the threshold, additional improvement is obtained for smaller SNR (stronger noise), which means that we effectively suppress the system with worse performance in noise. With the optimal threshold value $th = 2.5$ the combination achieves significant improvement for SNR>5. For stronger noises, the performance is only slightly inferior to the basic TRAP system.

The system combination with vector concatenation was a big surprise of our experiments. It achieves better performance on strong noises than the inverse entropy multi-stream combination and yet it keeps very good performance for week noises. Even larger improvement was observed for system combination where 2-dimensional time-frequency operator G2 [10] was used. This combination clearly outperformed all other systems.

We conclude, that the multi-band TRAP techniques and dimensionality reduction of TRAP vector by PCA (or DCT) are generally not good for recognition of noisy speech. It is due the inherent spreading of noise samples to larger area. The multi-stream combination techniques, namely the inverse entropy combination, improve significantly recognition of speech with weak noise (SNR $> 5$). The vector concatenation technique can bring improvement also in strong noises up to SNR $= 0$.

## References


1. B. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks," in *Proc. ICSLP 2004*, Jeju Island, KR, Oct. 2004.
2. P. Jain, H. Hermansky, and B. Kingsbury, "Distributed speech recognition using noise-robust MFCC and TRAPS-estimated manner features," in *Proc. of ICSLP 2002*, Denver, Colorado, USA, 2002.
3. A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI features for ASR," in *Proc. ICSLP 2002*, Denver, Colorado, USA, 2002.
4. P. Jain and H. Hermansky, "Beyond a single critical-band in TRAP based ASR," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003, pp. 437–440.
5. F. Grézl and H. Hermansky, "Local averaging and differentiating of spectral plane for TRAP-based ASR," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003.
6. F. Grézl, "Combinations of TRAP-based systems," in *Proc. TSD 2004*, Brno, Czech Republic, Sept. 2004, pp. 323–330.
7. D. Pearce, "Enabling new speech driven servicesfor mobile devices: An overview of the ETSIstandards activities for distributed speech recognition front-ends," in *Applied Voice Input/Output Society Conference (AVIOS2000)*, San Jose, CA, May 2000.
8. Cole R., Noel M., Lander T., and Durham T., "New telephone speech corpora at CSLU," in *Proc. of EUROSPEECH 1995*, Madrid, Spain, 1995, pp. 821–824.
9. H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream asr," in *Proc. ICASSP 2003*, Hong Kong, China, 2003.
10. F. Grézl, "Local time-frequency operators in TRAPs for speech recognition," in *Proc. TSD 2003*, Ceske Budejovice, Czech Republic, Sept. 2003, pp. 269–274.