Combinations of TRAP based systems *

František Grézl[1,2]

[1] Brno University of Technology, Faculty of Information Technology, Božetěchova 2, 612 66 Brno, Czech Republic, grezl@fit.vutbr.cz
[2] IDIAP , Martigny, Switzerland, Rue du Simplon 4, Case Postale 592, CH-1920
grezl.frantisek@idiap.ch

Abstract. We are introducing several methods for combination of systems based on temporal trajectories feature-level combination. Experiments were done to evaluate combination methods. Our results show improvement of recognition accuracy for combination of systems.

1 Introduction

In most cases, for solving a recognition task, one system with best performance is used. But good combination of two (or more) systems with poorer performance can give further improvement in accuracy. This paper shows the possibility of combination of TRAP based systems.

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 narrow frequency band. The nonlinear transformations — neural nets — are used for computing TRAP features.

Our previous studies and experiments indicate that information extracted from several (up to three) neighboring bands improves performance of the TRAP system [1]. Closer studies suggested that a simple pre-processing of a critical-band spectrogram (CRBS) prior to the cosine transformation and the TRAP feature extraction may be beneficial. This can be seen as additional feature stream. We are using some of known multi-stream combination techniques [2] to combine this features. But for TRAP system particularly it is possible to use different ways of combination and that is the main focus of this paper.

2 Trap Feature Extraction

After speech segmentation into 25 ms frames and computing of the power spectrum, spectrum energies are integrated into `M` filter bands (15 Bark scaled trapezoidal filters) and logarithm is taken. In each band, actual frame with +/− 50

* This research has been partially supported by Grant Agency of Czech Republic under project No. 102/02/0124 and by EC project Multi-modal meeting manager (M4), No. IST-2001-34485.

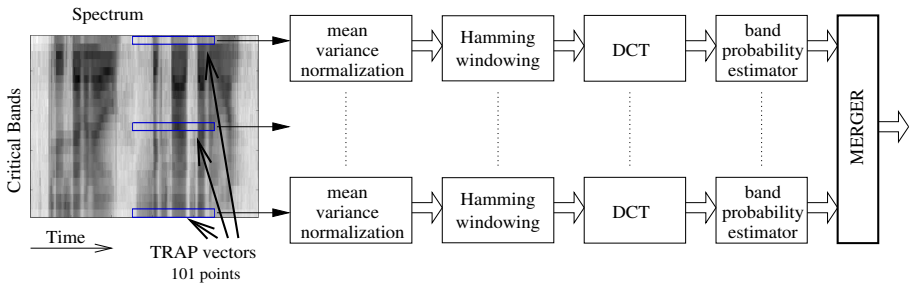

Fig. 1. TRAP system

G2			
	1	2	1
	0	0	0
	-1	-2	-1

Table 1. Coefficients of G2 operator

frames context is taken, so we have 101 points long TRAP vector. Mean and variance normalization of TRAP vectors follows. Finally Hamming windowing and discrete cosine transform is done.

The vector at the end of this processing is put into the **band probability estimator** — a three layer neural net. This net is trained to classify the input vector into one of the `N` classes. The input layer size is equal to the size of input vector, one hidden layer and the output layer, the size of which is equal to number of classes `N` (we used 29 phoneme classes). All output vectors are concatenated into a vector `M×N` points long. This vector goes through negative logarithmic nonlinearity and then forms the input for the **merger probability estimator**. Merger probability estimator is also a three layer neural net trained to classify input vector into the classes — the same target classes as the band probability estimators. The first layer has `M×N` points and the third layer has again `N` points. Its function is to merge particular band estimations into one final posterior probability vector. The scheme of the TRAP system is shown in Fig. 1.

Negative logarithm is taken and decorrelation using PCA is done on output of the merger probability estimator. This vector creates an input vector for standard GMM-HMM recognizer.

3 Critical Band Spectrogram Modification

The modifier operator (MO) was two-dimensional 3×3 operators known as Sobel filters in image processing [3]. Coefficients of the chosen MO are in Tab. 1.

We compute the new — modified — critical band spectrum as projection of the operator on the original spectrum. This operation is equivalent to the standard 2D FIR filtering. One point of modified CRBS (MCRBS) in given time

t and in given frequency band f, $MCRBS(t, f)$ is computed as

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

where f_c is the frequency context of the operator and t_c is its time context.

The processing of the MCRBS is same as for normal TRAP system. But the MCRBS size is going to be smaller by one point on each side of MCRBS. The missing points are repeated from previous (following) point in time domain. Missing frequency bands didn't cause any problem on the input of band probability estimator. We just have less of these estimators and consequently also smaller input vector to merger probability estimator.

4 System Combinations

4.1 Multi-stream combinations

The outputs from both systems are posterior probabilities and both systems have the same targets. The widely used technique is to average the output probabilities for the same class or to average the the log probabilities. The new approach of combination is based on the entropy of the system. We used Inverse entropy weighting with static threshold as described in [4] in our experiment. We were testing all these approaches.

4.2 Combination of band probability estimators outputs

The idea of this experiment was following: if we have the estimations of class probabilities on the output of band probability estimators, we can combine these probabilities. Now each TRAP sub-bands can be seen as multi-stream and we can combine outputs from band probability estimators.

First we tried to combine these probabilities directly in the merger probability estimator. This method is very simple and doesn't require any additional processing. The disadvantage is the size of the input vector, which almost doubles the size of vector in base TRAP system.

We added the pre-combination matrix to pre-process the outputs of band estimators and to form the input vector for merger probability estimator. The main task of pre-combination matrix was to capture dependencies in band probability estimator outputs and also to reduce the number of merger probability estimator inputs. We, of course, combine outputs belonging to the same class — **class vector**. Thus the pre-combination matrix can be seen as set of independent matrices, each one for one output class. The diagram is shown in Fig. 2. The pre-combined class vector (**pcv**) is computed as multiplication of class vector (**cv**) and its class pre-combination matrix (**CPM**) $pcv_n = cv_n \times CPM_n$ where $n = 1 \ldots N$ is the index of class. The input vector for merger probability estimator is created by concatenation of all pre-combined class vectors.

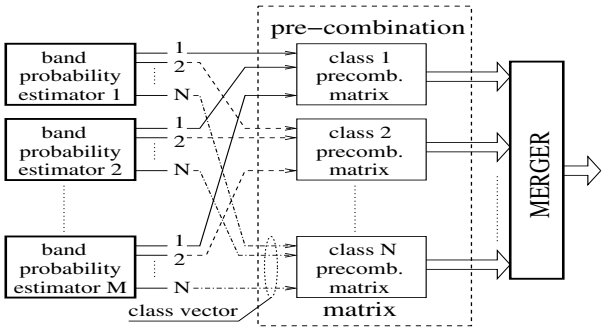

Fig. 2. Pre-processing of band estimators output by pre-combination matrix

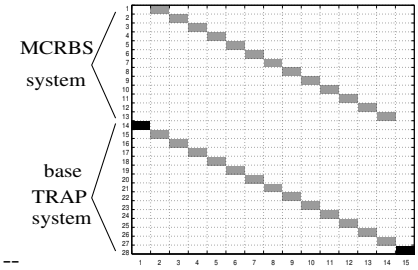

Fig. 3. System averaging matrix

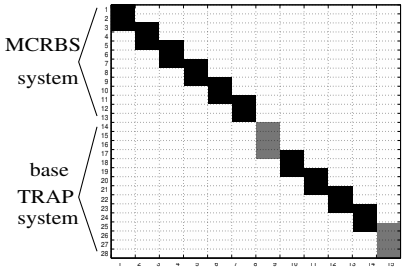

Fig. 4. Band averaging matrix

Several approaches were tested here. First we conducted an experiment where outputs which belong to the same frequency band and different system were averaged. This approach should capture the dependencies across the systems in the same frequency band. We called this *system averaging*.
In contrary of previous approach in the next experiment we average neighboring bands — *band averaging*. We suppose that the information from neighbor frequency bands from one system will be similar and can be merged.
The examples of matrices for these approaches are shown in figures Fig. 3 and Fig. 4

We also tried to emphasize more reliable outputs of band estimators in our experiments. For this, we performed the following analysis of the output of each band estimator:

- Pass the data used for merger probability estimator training through the band probability estimator and store the probability vector on the output of each band estimator with corresponding label.
- Count how many times the class with the highest probability is equal to the label. Divide this number by number of occurrences of given label. We will call this *hard hit vector*.
- Add together the estimated probabilities of the same class as label. Divide this number by number of occurrences of given label. We will call this *soft hit vector*.

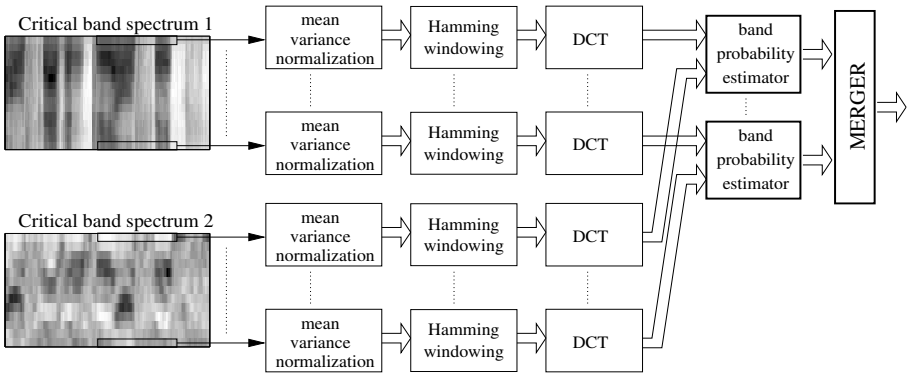

Fig. 5. Block diagram of system with vector concatenation

Each coefficient of the matrix is weighted by the corresponding hit vector coefficient. Then the matrix coefficients are "normalized" so that the sum of coefficients in each column of the pre-combination matrix is one.

The extreme way of weighting is to take just the "better system" output. The coefficient for system which performs better in given frequency band and for given class is one and the coefficient for the other system is zero.

We also performed an experiment where the pre-combination matrix was derived using Principal Component Analysis (PCA). PCA rotates the feature space in the direction of largest variability. We performed the PCA on each class vector separately so there is no mismatch between classes and the PCA assumptions are satisfied. We let the analysis technique decide which direction is important and rotate the feature space. We reduced the dimensionality of each class-vector to fifteen points. The PCA base vector are in columns of the pre-combination matrix.

4.3 Vector concatenation

Vectors from different TRAP systems are concatenated on the input of band probability estimator.

Since there is different number of bands in modified and original CRBS, we have to take these missing vectors from the closest frequency MCRBS. Thus we have the same number of critical bands now.

All processing (normalization, windowing, DCT) is done on each vector independently and finally, vectors are concatenated into one. The block diagram in Fig. 5 shows the processing for system with vector concatenation.

5 Experimental setup

The task is recognition of eleven words (digits). The test set was derived from CSLU Speech Corpus [5]. This part contains utterances only with connected

digits. There are 2169 utterances with total length about 1.7 hours. There are 12437 words in this set. The recognizer is HTK based GMM-HMM system. Each word is modeled by sequence of context independent phoneme models. These consist of five states, three mixture components per state. The training set contains 2547 utterances with total length about 1.2 hours. This set is also derived from the CSLU Speech Corpus and utterances containing only connected digits are used. A subset of CSLU Speech Corpus was used also for training the merger probability estimator. This set contains 3590 utterances with total length of about 1.8 hours. No restrictions apply to this set. A subset of OGI Stories database [6] was used for training the band probability estimators. This set contains 208 utterances with total length of about 2.7 hours.

The number of target phonemes for training probability estimators is $\mathtt{N} = 29$. Target phonemes are these which occur in digits utterances. Others phonemes are not used for training but they create context in TRAP vectors. 50 cosine basis were used in DCT. The size of the band probability estimator's input layer is 50 points when single operator is used and 100 points when there is a vector concatenation. Speech spectrum was integrated into $\mathtt{M} = 15$ Bark-scaled trapezoidal filters. There are 13 bands when frequency operator is used. All neural nets used in probability estimators have 300 units in hidden layer.

6 Results

The performance measurement for tested systems was word error rate (WER) computed as:

$$WER = 1 - \frac{hits - insertions}{number\ of\ words} \times 100\%.$$

We also calculated that one system is significantly better than another if the difference in theirs WER is at least 0.5% for our test set. This was calculated on confidence level 95%. If we want to be 99% confident, the difference in WER has to be at least 0.7%.

Results for basic TRAP system and G2 modified TRAP system are shown in Table 2. The results for system combinations are shown in Table 3.

7 Conclusion And Discussion

The results show that even though the MTRAP system itself did not achieve the performance of basic TRAP system, their combination brings significant improvement. It means there is complementary information which can be used to reduce the system WER.

The multi-stream methods show the same tendencies as in [2]. We got the poorest performance for simple averaging of the output probabilities. The inverse entropy gains a slight improvement. The advantages of inverse entropy criteria, as was shown in [4] is mainly for noisy speech which is not our case. Averaging the log probabilities gives better performance for multi-stream approach.

system	WER [%]
basic TRAP	6.1
G2 MTRAP	7.1

Table 2. WER [%] of basic TRAP and G2 MTRAP

system	WER [%]
multi-stream average	4.8
multi-stream log average	4.5
multi-stream inverse entropy	4.6
no pre-combination	4.8
averaging - system	4.8
averaging - bands	4.6
W-h averaging - system	4.9
W-s averaging - system	4.6
W-h averaging - bands	4.6
W-s averaging - bands	4.5
PCA	4.4
better system	4.5
vector concatenation	4.2

W-h — weighting according to hard hit vector;
W-s — weighting according to soft hit vector

Table 3. WER [%] of systems combinations

System combination based on combination of band probability estimators outputs gives us comparable results. This combination leaves out one merger probability estimator so the total number of weights in the system is smaller.

System averaging pre-combination of class vectors preserves the information derived from narrow frequency bands and combines the information obtained using different systems. Band averaging pre-combination of class vector averages the information over several frequency bands and keeps information from different systems separated. Designing of different band combination is also possible. We could also design pre-combination which will combine both informations — from different systems and from different bands, but this approach has not been tested. However the results are very similar and the outcome of such combination will be further reduction of number of weights with the same performance.

The weighted average and "Better system" pre-combination brings into account the performance of particular probability estimator. But in our case the data used for band estimator training are different from data used for final test and neither the performance nor the weights are optimal for the final data.

The PCA pre-combination performs very well. For this we need the data analysis but it is possible to keep only several most important points in the output vector.

Combination method	no. of weights
multi-stream	1448 k
pre-combination	1309 k
vector concatenation	1178 k

Table 4. Number of weights in TRAP system for different combination methods

The vector concatenation is an easy method which gives us the best results. This method also decreases the number of probability estimators, they have only bigger number of inputs. Responsibility of choosing the proper information is left on the band estimator.

The Table 4 gives the number of weights in the TRAP systems. It is obvious that the system with vector concatenation which also has the best performance, has also the lowest number of weights.

We show the advantages of combination of TRAP system based on different critical band spectrograms. Several approaches of combination of TRAP based system were tested. Overall the results are similar, but the difference between the poorest and the best performance is still above the confidence level. Furthermore, our results show, that the least complicated system in terms of used weights has the best performance.

Combination of more systems is also possible using introduced methods. This combination is happening on feature level so there is no need for any changes in the recognition system. Although presented task was simple, our current experiment show advantages also for much bigger and complicated task such is the large vocabulary continuous speech recognition.

References

1. P. Jain and H. Hermansky, "Beyond a single critical-band in trap based asr," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003, number ISSN 1018-4074.
2. Q. Zhu A. Stolcke N. Morgan, B. Y. Chen, "Trapping conversational speech: Extending tarp/tandem approaches to conversational telephone speech recognition," in *Proc. ICASSP 2004*, Montreal, Canada, 2004.
3. Anil K. Jain, *Fundamentals of digital image processing*, Number ISBN 0133361659. Prentice Hall, 1988.
4. H. Bourlard H.Misra and V. Tyagi, "New entropy based combination rules in hmm/ann multi-stream asr," in *Proc. ICASSP 2003*, Hong Kong, China, 2003.
5. Noel M. Cole R., Fanty M. and Lander T., "Telephone speech corpus development at CSLU," in *Proc. of ISCLP 1994*, Yokohama, Japan, 1994, pp. 1815–1818.
6. Lander T. Cole R., Noel and Durham T., "New telephone speech corpora at CSLU," in *Proc. of EUROSPEECH 1995*, Madrid, Spain, 1995, pp. 821–824.