TRAP based features for LVCSR of meeting data

František Grézl [1,2], *Martin Karafiát* [1], *Jan Černocký*, [1]

[1] Brno University of Technology, Faculty of Information Technology,
Božetěchova 2, 612 66 Brno, Czech Republic,
{grezl,karafiat}@fit.vutbr.cz

[2] IDIAP , Martigny, Switzerland,
Rue du Simplon 4, Case Postale 592, CH-1920
grezl.frantisek@idiap.ch

Abstract

This paper describes using temporal patterns (TRAPs) feature extraction in large vocabulary continuous speech recognition (LVCSR) of meeting data. Frequency differentiation and local operators are applied to critical-band speech spectrum. Tests are performed with HMM recognizer on ICSI meetings database. We show that TRAP features in with standard ones lead to improvement of word-error rate (WER).

1. Introduction

The novel TRAP based features are used in automatic speech recognition (ASR) for several years. So far these features were used with combination with standard ones and mainly for simple tasks with small vocabulary [1, 2]. But our last research also led to significant improvement on small vocabulary task using just TRAP based features [3].

These results encouraged us to use TRAP features in more complex task, large vocabulary continuous speech recognition (LVCSR) of meeting data. It introduced TRAPs into new difficulties such as cross-talks, vocal noises, unfinished words, non-native speakers, etc.

The TRAP features are based on narrow band spectrum with long time context. Vectors which actually evolution of energy in critical bands – are processed independently. This is the basic idea of multi-band processing described also in [4]. The further improvement of TRAP based features was achieved by adding another feature stream. The new feature vector is derived in the same way as conventional TRAP vector but it is derived from critical band spectrogram (CRBS) which was modified by a local operator performing band differentiation. This modification was proposed as outcome of previous studies and experiments [5] which shows the importance of differentiation information for speech recognition based on TRAP features.

The original TRAP vector and vector derived from modified critical band spectrogram (MCRBS) are then concatenated and fed into a multi-layer perceptron (MLP) band probability estimator. The partial estimations from all bands are transformed into final estimation by another MLP merger estimator. The scheme of the our TRAP system is shown in Fig. 1.

Outputs from merger MLP are after decorrelation used as features — MCRBS TRAP features — for GMM-HMM system.

2. Feature extraction

This section reviews basic principles of the TRAP feature extraction, studied and applied in this work. The short-term critical band spectrum is computed in 10 ms (100 Hz) analysis steps using the same module as used for the critical band spectral estimation in PLP analysis [6] and the logarithm of the estimated critical band spectral densities is taken. Let us denote the number of critical bands as M.

This CRBS is modified by a one dimensional local operator which does the band differentiation – BD or by a two dimensional Gabor operator known from image processing as a edge detector – G2. The coefficients of these operators are shown in table 1. We compute the MCRBS as projection of the modifying operator (MO) on the original spectrum. One point of modified MCRBS in given time t and in given frequency band f, $MCRBS(t, f)$ is computed as

BD

-1
0
1

G2

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

Table 1: Spectrum modifying local operators

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

where f_c is frequency context of the operator and t_c is its time context. When one dimensional operator is used, the context in other direction will be zero.

Then for each critical band from both spectrograms the following process was performed :

- mean normalization,
- frame with context between ±20 to ±50 is taken,
- Hamming window is applied emphasizing the center of the vector,
- the vector is transformed by a discrete cosine transform (DCT) reducing the size of the vector to a half.

The vectors from different CRBS and for corresponding bands are then concatenated to create input for band probability estimator MLP. Due to modifying process the resulting MCRBS has reduced number of bands. The first/last band in MCRBS is therefore

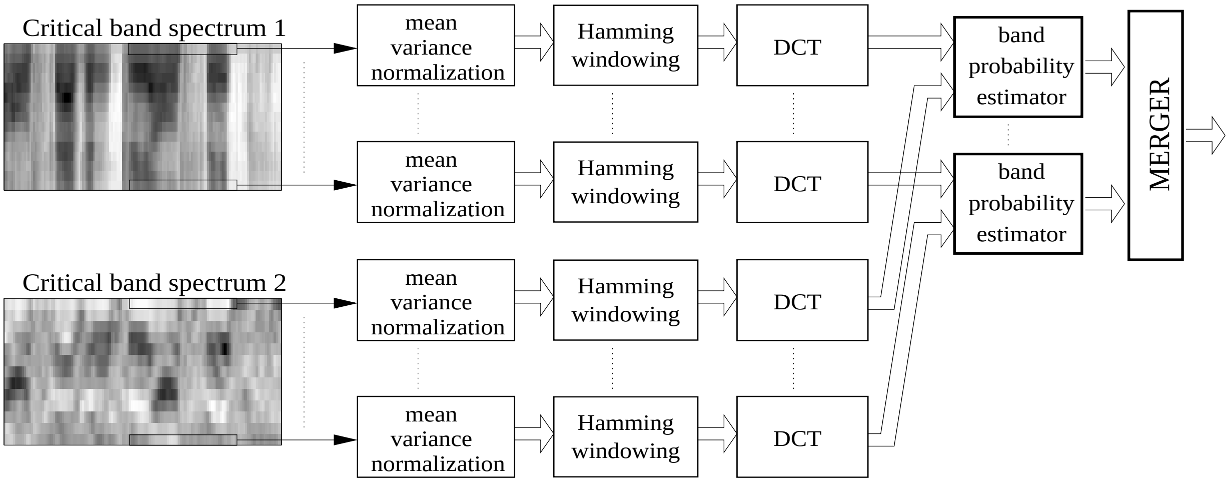

Figure 1: Block diagram of feature extraction process

repeated from consecutive/preceding band, so we obtain the same number of critical bands as in original CRBS.

The band probability estimator — a three layer neural 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.

All output vectors from all band estimators are concatenated into a vector $M \times N$ 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 $M \times N$ 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.

The output probabilities are gaussianized by logarithm and decorrelated using PCA. This vector creates an input vector for standard GMM-HMM recognizer

3. GMM-HMM recognition system

Context independent one gaussian phoneme HMMs are directly initialized and reestimated on trainig data with associated phoneme level transcriptions.We use 3 states models for all phonemes and noises. The loop from end state to start one is added in noise models.

These models are used for initialization of context dependent ones. They are copied out according to all occurrences of triphones in training data and reestimated. A generation of tied states and extension for unseen triphones by decision tree method follow. Number of tied states is adjusted into the range 3000-3500. The output 8 gaussian mixture triphone models are generated by splitting mixtures and iterative Baum-Welsh reestimations.

We test the features on a simple recognizer based on word internal triphones and trigram language model. A generation of output ASR transcriptions is performed by fast stack decoder which is more suitable for large vocabulary task than standard Viterbi decoder. According to our preliminary results, the recognition accuracy of the Viterbi search is slightly higher but stack decoder is significantly less time and memory consuming. Strong pruning is used for further performance acceleration with loosing 2% accuracy only. This decreased time needed for decoding from 150× 8× real time only.

4. Experiment description

4.1. Database

We use the ICSI meetings database [7] for the experiments. It contains the recordings of real spontaneous meetings with cross-talks, unfinished words, background speech and all kinds of speaker noises. The speakers in the database are often non-native. All these conditions makes the recognition very difficult.

Two sets of data were derived:

Training part consist of 39.4 hours of speech from 26 speakers (4 female, 22 male, 12 native, 14 non-native).

Test part consist of 1 hour of speech randomly selected from three meetings. There is 7 speakers (2 female, 5 male, 4 native, 3 non-native).

The data parts were derived so that no speaker occurs in both parts. Only parts of signals with speech activity were taken. There is 10282 words of which 113 only appears only in the test part.

4.2. TRAP processing

The number of critical bands used in PLP analysis was `M = 19`.

We optimized the context length and the normalization style. The considered contexts were ±20, ±30, ±40, ±50 frames. The tested normalizations kinds were no normalization, mean normalization on the utterance, mean normalization on TRAP vector. The optimum was found for context ±20 frames and utterance based normalization.

Both band estimator and merger estimator MLPs were trained against the hard targets to estimate the probability of $N = 45$ classes. The 40 classes out of 45 were “proper” phonemes, 2 classes were “hesitation” phonemes, 2 classes were noise kinds, and 1 class was silence.

The size of band probability estimators was 20 inputs, 100 hidden units and 45 outputs.

The merger MLP input layer was quite big in our case. It contained `M×N = 19×45 = 855` units in the input layer. The size of hidden layer was 300 units for the testing purposes. After the optimal normalization and time span was found the number of hidden units was increased to 1500. The number of outputs was 45.

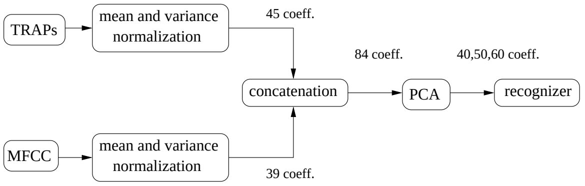

Figure 2: Concatenation and reduction of features by PCA.

The MLPs were trained only on 10 hours of train data (17700 files).

4.3. Language model and dictionary

Data for training of our language model were taken from transcriptions of train part ICSI database (53099 sentences) and whole Switchboard one (248581 sentences). Each sentence from ICSI database was copied by for better balancing of training data amounts and decreasing of perplexity. The perplexity was evaluated on Switchboard transcriptions and ICSI training data trascriptions (see in Tab. 2).

Data balance	Perplexity
1xICSI 1xSWB (no balance)	127.74
2xICSI 1xSWB	122.46
3xICSI 1xSWB	119.53
4xICSI 1xSWB	119.48
5xICSI 1xSWB	116.74
6xICSI 1xSWB	117.32

Table 2: Results from perplexity optimization

The dictionary was created by merging the ICSI meetings dictionary with Switchboard dictionary. There was totally 36136 words.

4.4. MFCC baseline features

The MFCC features were derived using HTK tools. Twelve cepstral features were computed appended with zeroth cepstral coefficient. The first and second derivatives were computed over 5 frames and added to the feature vector. The feature vectors were normalized over the utterance.

4.5. MFCC TRAP feature combination

The combination of MFCC and MCRBS TRAP features was also tested. The MFCC coefficients with theirs derivatives were first normalized over all data and then the MCRBS TRAP based features (after logarithmic transformation and PCA decorrelation) were appended to them. PCA was performed to reduce the dimensionality of the feature vector. MFCC - MCRBS TRAP feature vectors with 40, 50 and 60 coefficients were tested. This is displayed in Fig. 2.

4.6. Used tools

The recognition system uses the combination of packages: the HTK toolkit [8] for a parametrization of input speech, training and working with acoustic models ;the Du-coder [9] - a stack decoder; and the SRILM toolkit [10] - a language model training tool.

5. Recognition results

The results for MCRBS TRAP features are shown on the table Tab. 3.

Explanation of abbreviations:

- G2 MCRBS TRAP - The MCRBS TRAP features used modifying operator G2
- BD MCRBS TRAP - The MCRBS TRAP features with BD modifying operator
- MFCC-TRAP No. - Combination of MFCC baseline features and BD MCRBS TRAP features. The number following of the name gives the information about feature vector size after PCA.

However the performance of the G2 MCRBS TRAP system seems to be poorer than the MFCC baseline, other systems are more promissed. The BD MCRBS TRAP is giving same accuracy as MFCC but merging this features with baseline ones and dimensionality reduction by PCA is giving nice improvement of accuracy. Optimal number of output feature vector size was found for 40 coefficients and it gives 1.7% relative word error improvement.

features	GMM-HMM
MFCC baseline	46.4
G2 MCRBS TRAP	46.1
BD MCRBS TRAP	46.4
MFCC-TRAP 40	47.2
MFCC-TRAP 50	46.5
MFCC-TRAP 60	47.0

Table 3: Recognitions results WER [%].

6. Conclusion and discussion

First experiments were aimed at the comparison of TRAP-based features with the MFCC baseline. The first three rows in Table 3 show, that TRAPs obtained using the G2 operator perform worse than the baseline. On the other hand, using band differentiation (BD) only provides the same result as MFCCs. While in previous works [3] we have seen G2 operators outperforming or at least provide similar results to BD, here we have a hit. We might explain this behavior by the use of context-dependent models (CD-HMM) in this work - the G2 operator performs integration in time, and CD-HMMs need finer temporal information.

In the following experiments, the merging of MFCC and TRAP features was tested. As the resulting feature vector would become huge, PCA was used to reduce its dimensionality. We have observed differences while varying the number of target coefficients after dimensionality reduction — the optimum was found at 40, where a nice 1.7% relative error improvement was observed compared to the MFCC baseline. We can conclude that TRAP features, by bringing additional infomation from long time spans, are beneficial for meeting data recognition.

We are currently testing also more complex recognition system based on lattice generation, their rescoring by more accurate crossword triphones and speaker adaptation. For MFCCs, this system increases accuracy by about 4%. It must be verified, if we will see similar improvement also for TRAP features or their combination with MFCCs.

Another way to improve our system is to use Heteroscedastic Linear Discriminant Analysis [11] on concatenated feature vectors instead of PCA. In the data, this transform is looking for dimensions with maximum discriminability instead of vatiability and on contrary to LDA it imposes less assumptions on the data.

Concerning the recognition part of work, our aims are to exploit differences of meeting data from classical LVCSR (availability of multiple channels, side information, speaker ID) to further improve the recognition results.

7. Acknowledgments

This research has been partially supported by EC projects Multimodal meeting manager (M4), No. IST-2001-34485 and Augmented Multi-party Interaction (AMI), No. 506811 and by Grant Agency of Czech Republic under project No. 102/02/0124. Jan Černocký was supported by post-doctoral grant of Grant Agency of Czech Republic No. GA102/02/D108.

8. References

[1] P. Jain, H. Hermansky, and B. Kingsbury, “Distributed speech recognition using noise-robust MFCC and TRAPS-estimated manner features,” in *Proc. ICSLP 2002*, (Denver, Colorado, USA), 2002.

[2] B. CHan, S. Chang, and S. Sivadas, “Learning discriminative temporal patterns in speech: Development of novel traps-like classifiers,” in *Proc. Eurospeech 2003*, no. ISSN 1018-4074, (Geneva, Switzerland), 2003.

[3] F. Grézl and H. Hermansky, “Local averaging and differentiating of spectral plane for TRAP-based ASR,” in *Proc. Eurospeech 2003*, no. ISSN 1018-4074, (Geneva, Switzerland), 2003.

[4] S. R. Sharma, *Multi-stream approach to robust speech recognition*. PhD thesis, Oregon Graduate Institute of Science and Technology, Oct. 1999.

[5] P. Jain and H. Hermansky, “Beyond a single critical-band in trap based asr,” in *Proc. Eurospeech 2003*, no. ISSN 1018-4074, (Geneva, Switzerland), 2003.

[6] H. Hermansky, “Perceptual linear predictive (PLP) analysis for the speech,” *J. Acous. Soc. Am.*, pp. 1738–1752, 1990.

[7] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters.(2003), “The ICSI meeting corpus,” in *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal processing (ICASSP)*, (Hong Kong), 2003.

[8] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, *The HTK book*. Cambridge, UK: Entropics Cambridge Research Lab., 2002.

[9] D. Willett, C. Neukirchen, and G. Rigol, “DUCODER-the duisburg university LVSCR stackdecoder,” in *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal processing (ICASSP)*, (Istanbul), 2000.

[10] A. Stolcke, “SRILM - an extensible language modeling toolkit,” in *Proc. International Conference on Spoken Language Processing (ICSLP)*, (Denver), pp. 901–904, 2002.

[11] L. Burget, *Speech Recognition System Complementarity and System Combination*. PhD thesis, Brno University of Technology, in progress 2004.