Spectral Plane Investigation for Probabilistic Features for ASR

M-D-17

František Grézl [1,2]

[1]Brno University of Technology, Faculty of Information Technology, Božetěchova 2, 612 66 Brno, Czech Republic
e-mail: grezl@fit.vutbr.cz

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

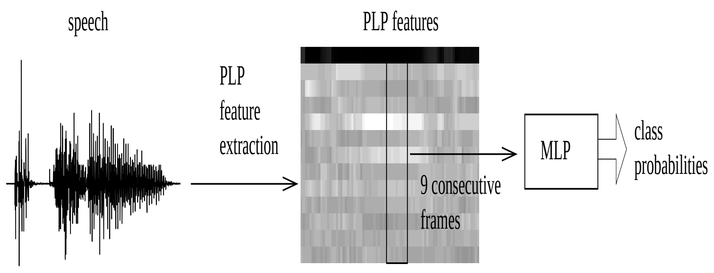
Abstract

This poster discusses effect of different frequency and time context of input critical band spectrogram block in two stage TRAP-TANDEM feature extraction. Best performance was observed when splitting input spectrogram into rather narrow frequency slices with long time context.

1 Probabilistic features

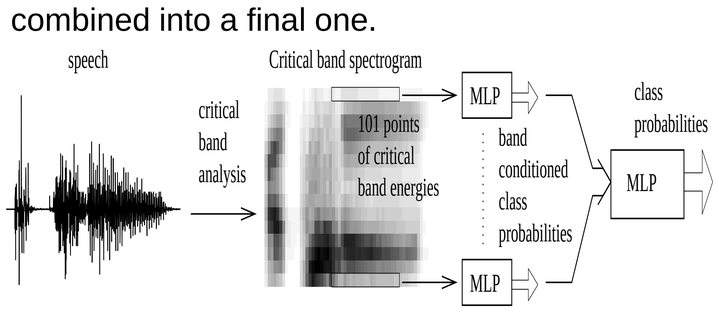
- Probabilistic features are class probabilities transformed to the form suitable for the GMM/HMM recognizer.
- Class probabilities are estimated by a nonlinear classifier – feed-forward multi layer perceptron (MLP).

TANDEM – estimates class probabilities from several consecutive frames of standard features.

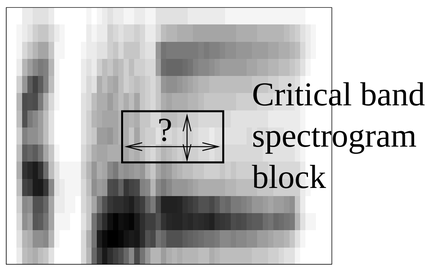


TRAP – estimates class probabilities from critical band spectrogram. This is two stage process:

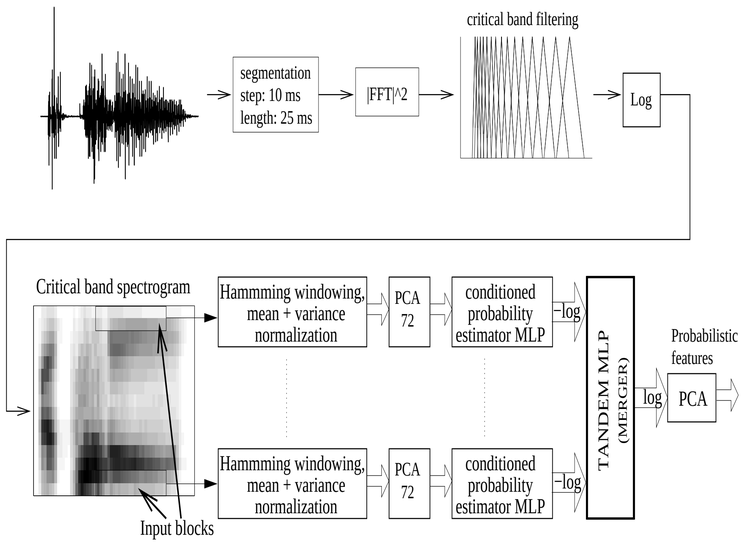
1. Class probabilities are estimated for each critical band.
2. Critical band conditioned class probabilities are combined into a final one.



- Optimal size of critical band spectrogram block for TRAP-TANDEM feature extraction – **?**



2 Computation of probabilistic features

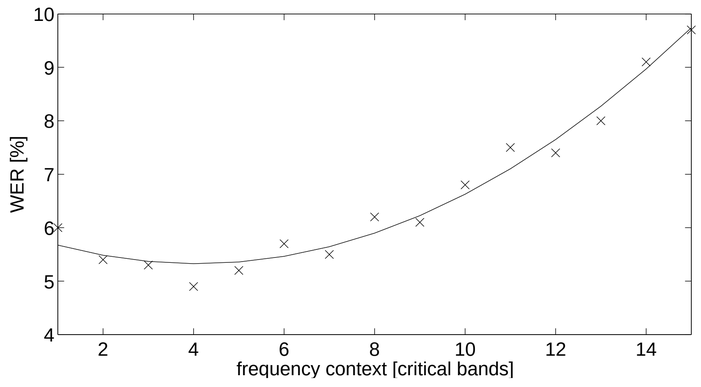


3 Experimental setup

- Recognition of eleven concatenated digits.
- Test on the testing part of OGI Numbers database.
- HTK based GMM-HMM recognition system.
- Context independent phoneme models used.
- HMMs are trained on a training part of OGI Numbers database
- A training part of OGI Numbers database was also used for training the merger probability estimator.
- A subset of OGI Stories database was used for training the band conditioned probability estimators.
- 29 phoneme targets (phonemes in numbers).

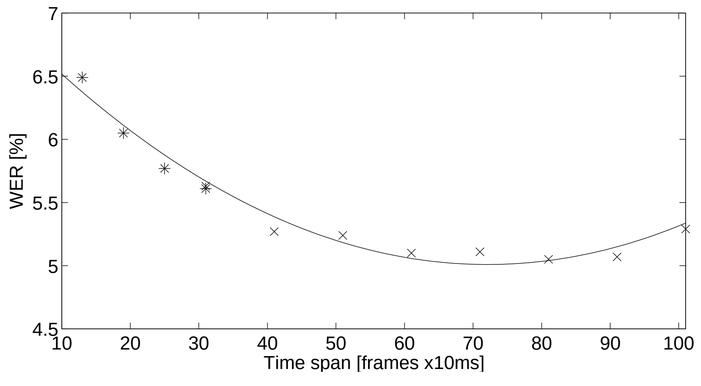
4 Changing frequency context

- Frequency context (K) vary from 1 to 15 bands
- Frequency shift between two neighboring critical band spectrogram blocks is 1 band.
- Number of band conditioned MLPs is $15 - K + 1$
- TRAP-TANDEM reduced to TANDEM when frequency context of critical band spectrogram block was 15 bands (no band conditioned estimators).
- Temporal context fixed at 101 frames.



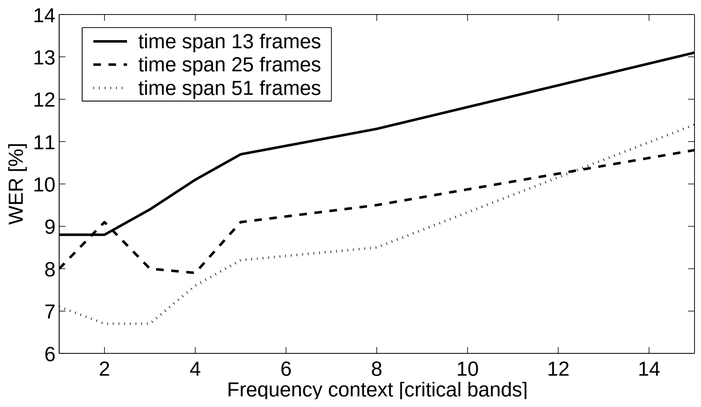
5 Changing time context

- Frequency context is fixed at 3 bands.
- Frequency shift between two neighboring critical band spectrogram blocks is 1 band.
- Temporal context 31 to 101 frames → PCA returns 72 elements.
- Temporal context 13 to 31 frames → PCA returns 36 elements.



6 Changing time and frequency context

- PCA was not used in probabilistic features computation.
- Critical band spectrogram blocks do not overlap.



7 Conclusion

- Word error rate has a wide and flat minimum for both time and frequency contexts.
- Best performance was observed when splitting the critical band spectrogram into blocks with narrow frequency context around 4 bands.
- Beneficial to use temporal context between 50 and 90 frames ($\sim 0.5 - 1$ second).

Acknowledgement

This work was partially supported by EC project Multimodal meeting manager (M4), No. IST-2001-34485, European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction) and Grant Agency of Czech Republic under project No. 102/05/0278.

References

[1] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP 2000*, Turkey, 2000.

[2] S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky, "Feature extraction using non-linear transformation for robust speech recognition on the Aurora database," in *Proc. ICASSP 2000*, Turkey, 2000.

[3] Sunil Sivadas, *Tandem Feature Extraction for Automatic Speech Recognition*, Ph.D. thesis, OGI School of Science & Engineering Oregon Health & Science University, Nov. 2004.

[4] S. R. Sharma, *Multi-stream approach to robust speech recognition*, Ph.D. 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*, Geneva, Switzerland, 2003, number ISSN 1018-4074.

[6] P. Schwarz, P. Matějka, and J. Černocký, "Recognition of phoneme strings using TRAP technique," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003, number ISSN 1018-4074.