Robust heteroscedastic linear discriminant analysis and LCRC posterior features in meeting data recognition

Martin Karafiát, František Grézl, Petr Schwarz, Lukáš Burget and Jan Černocký

Speech@FIT, Department of Computer Graphics and Multimedia
Faculty of Information Technology, Brno University of Technology

Plan

- Heteroscedastic Linear Discriminant analysis (HLDA)
  - Robust modification
  - Silence reduction
- Posterior features
  - PLP based posterior estimator - FeatureNet
  - Split context posterior estimator - LCRC
- Application on the CTS and meeting data
- Conclusion

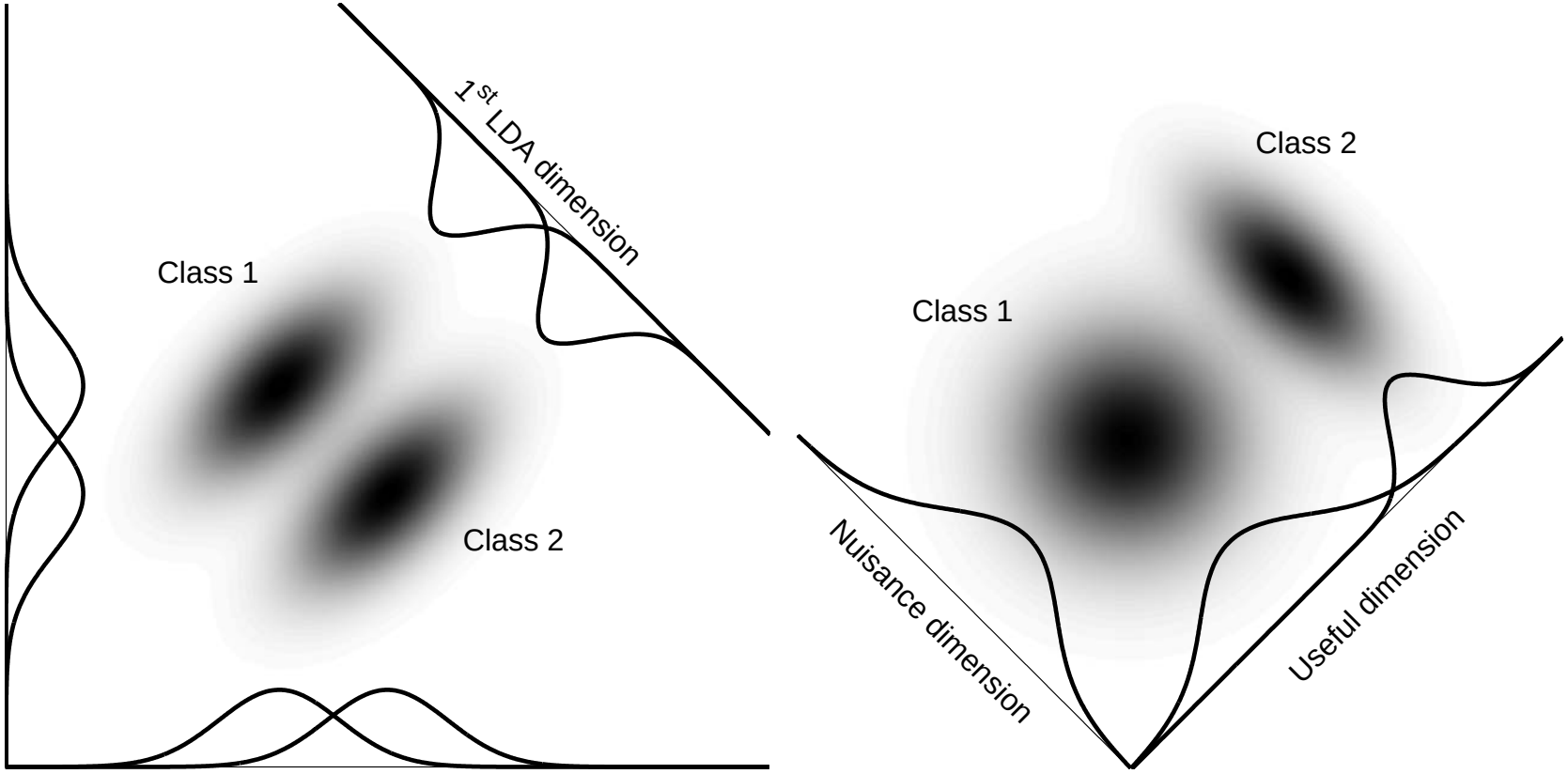
Goals

- Feature extraction - dimensionality reduction and feature de-correlation
  - de-correlation - diagonal covariance Gaussian modeling
  - dimensionality reduction - preserve discriminative information and remove nuisance dimension.
- adding posterior features - we're desperately searching for source of complementary information, if we find one at the feature level, it can bring nice improvement and cheaper than more training data, better LM or better (and slower!) decoder.

HLDA

- Feature transform - projection into space where features are decorrelated - useful for diagonal covariance modeling.
- Class based
  - training space has to be split into the classes.
  - each class has different covariance matrix.
  - if all class share the same covariance matrix - HLDA turns to well known LDA.
- Dimensionality reduction - find dimensions where feature classes are best separated.

HLDA vs. LDA


1st LDA dimension
Class 1
Class 2
Class 2
Class 1
Nuisance dimension
Useful dimension

Robust HLDA in HMM systems

Reliability of class covariance estimates:

- 1 class = 1 Gaussian.
- Clustering process is easily done by classic forward-backward algorithm.
- Some Gaussians can end up with insufficient amount of data to obtain proper full covariance estimates. Basic HMM system use variance flooring to solve this problem.
  But HLDA?

Solution:

- Smoothing by average within class covariance matrix $\mathbf{\Sigma}_{WC}$, which is weighted average of class full covariance matrices.
- We have developed two smoothed techniques.
  1. Smoothed HLDA
  2. MAP smoothed HLDA

Smoothed HLDA - SHLDA

- technique based on combination HLDA and LDA
- full covariance matrixes in **all classes** are smoothed by within class covariance matrix $\mathbf{\Sigma}_{WC}$.
- $\check{\mathbf{\Sigma}}_j = \alpha\hat{\mathbf{\Sigma}}_j + (1-\alpha)\mathbf{\Sigma}_{WC}$,
  where α is smoothing factor.

  $\alpha = 0 \quad \Rightarrow \quad SHLDA = LDA$

  $\alpha = 1 \quad \Rightarrow \quad SHLDA = HLDA$

MAP smoothed HLDA - MAP-SHLDA

- not important to smooth classes with enough training data.
- smoothing factor should depend on amount of training data for each particular class.
- we apply Maximum a posteriori (MAP) adaptation of covariance matrixes. $\mathbf{\Sigma}_{WC}$ is considered as a prior:

$$\check{\mathbf{\Sigma}}_j = \mathbf{\Sigma}_{WC}\frac{\tau}{\gamma_j + \tau} + \hat{\mathbf{\Sigma}}_j\frac{\gamma_j}{\gamma_j + \tau},$$

where τ is a control constant.

Silence reduction in HLDA - SR-HLDA

Silence is special class in speech recognition:

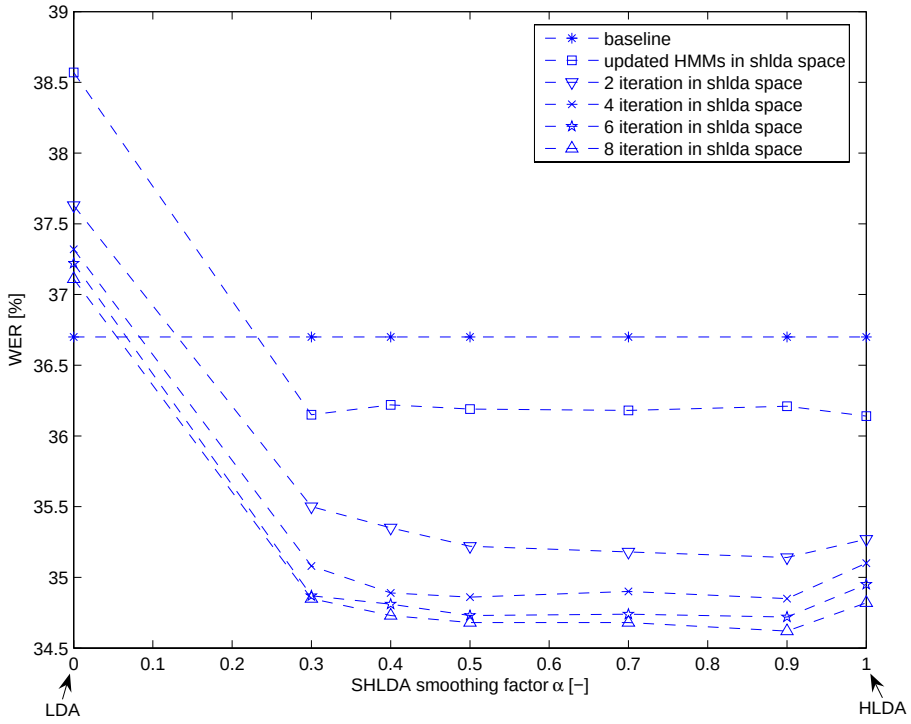
- significantly differs from other classes.
- is partly contained in phonemes like stops, plosives...
- much more data for silence class than for other classes.
- we prefer to find dimensions where *voice classes* will be the best separated (we don't want from HLDA to do voice/silence separation).
- **this big and less important class could have negative effect on HLDA performance!**

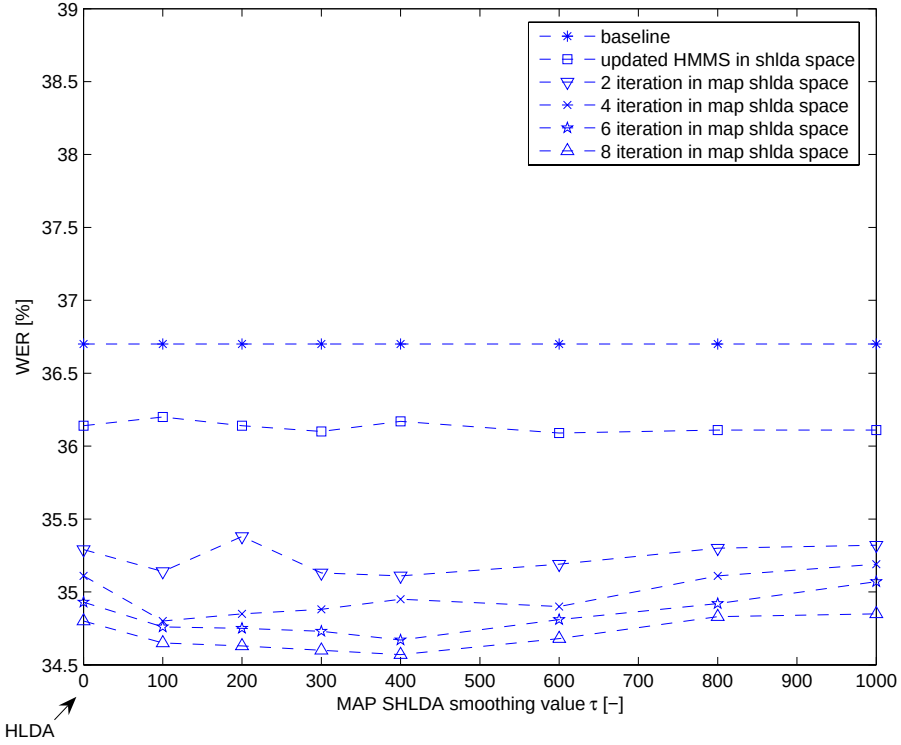
Solution:

- 2-HLDA system - significantly increases complexity of the system.
- Decrease influence of silence in HLDA computation - can be easily done by dividing occupation counts of silence classes by silence reduction constant (SR) - better than discarding frames.

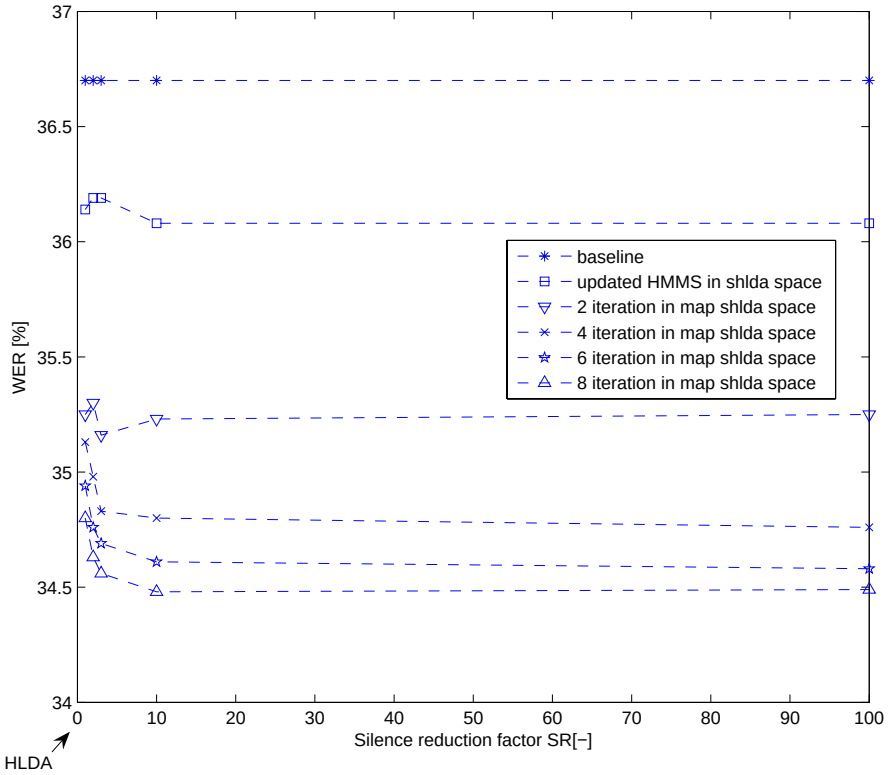
Comparison of HLDA systems

- Trained on 270h hours of conversation telephone speech (CTS) data.
- Tested on eval01 data.





Performance of silence reduced HLDA



Comparison of HLDA systems - Summary

System	WER [%]
Baseline (no HLDA)	36.7
HLDA	34.8
SHLDA	34.6
MAP-SHLDA	34.6
SR-HLDA	34.5

- **The SR-HLDA was fixed for next experiments with posterior features.**

POSTERIOR FEATURE SYSTEMS

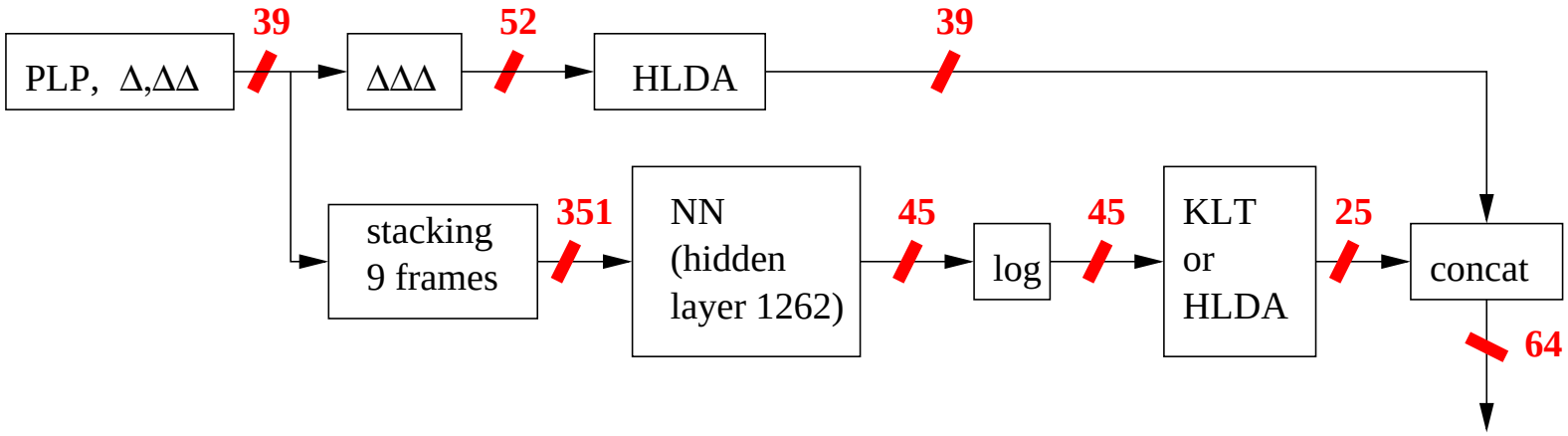
1. Gives complementary information to standard (MFCC or PLP) features.
2. Based on Neural Net (NN) posterior estimator.
3. A decorrelation and possible dimensinality reduction needed for further processing by HMM.

The final features are constructed as concatenation of two different feature streams.

1. PLP HLDA 39dimensional features.
2. Decorrelated NN outputs. We played with two posterior systems.
   - FeatureNet.
   - LCRC posterior estimator.

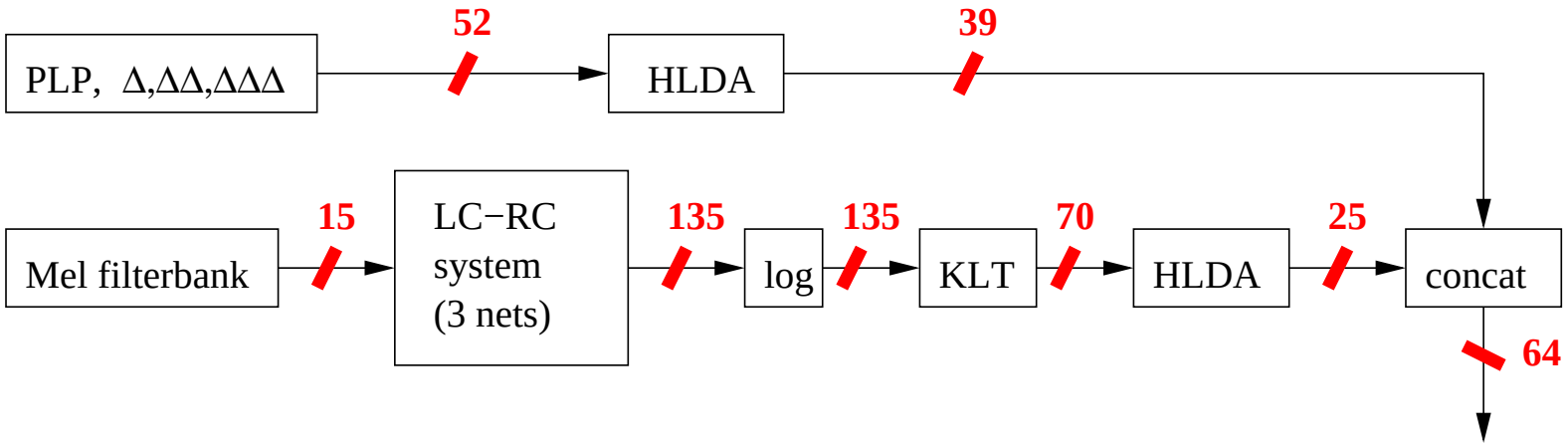
FeatureNet

- 1 Neural Net.
- Input - 9 PLP frames. Input dimensionality 39. It gives 9x39 feature vector.
- NN output - 45 vector size (45 phonemes).
- Decorrelation and dimensionality reduction to 25 final vector output.

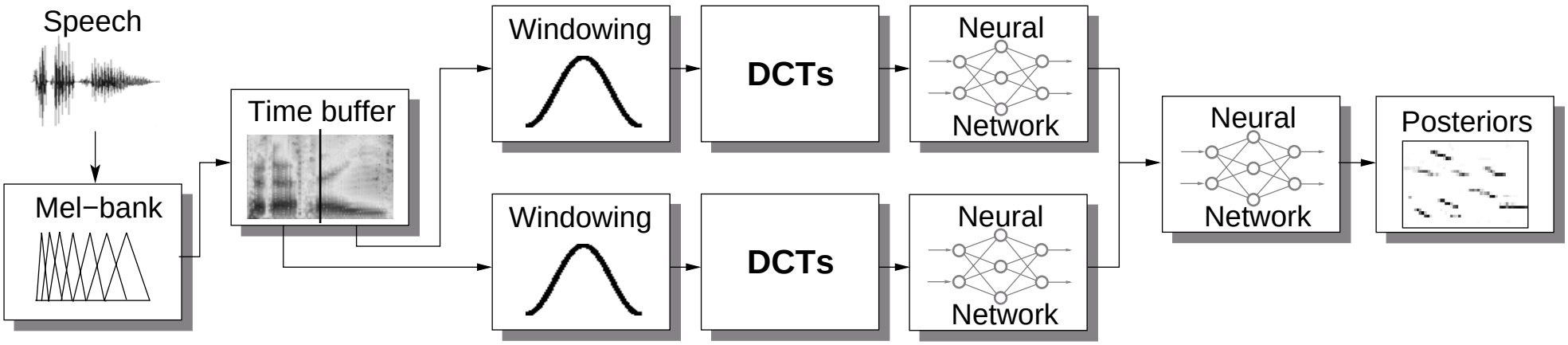


LCRC posterior system

- 3 Neural Nets (Left context, Right context, Merger)
- Input - 31 frames from Mel filter bank output. 15 used for left context NN and 15 for right context one. Middle frame is overlapped to the both sides.
- Output - 135 phoneme state output. 3state for each phoneme.
- Decorrelation and dimensionality reduction to 25 final vector output.



LCRC posterior estimation



Quite sophisticated structure... Why do we use scheme like this?

- The scheme was already tuned for phoneme recognition (the best result on the TIMIT).
- successfully used in other tasks as Language identification and phoneme based keyword spotting.

Results of posterior systems on CTS data

- HMM was trained on 270h hours of cts data (h5train03 cambridge training set).
- NN was trained on 10h subset HMM training.
- Tested on eval01 data.

System	WER [%]
PLP SR-HLDA	34.5
PLP SR-HLDA + PLP-posteriors KLT25	33.8
PLP SR-HLDA + PLP-posteriors HLDA25	33.3
PLP SR-HLDA + LCRC-posteriors HLDA25	32.6

Posterior system on meeting data (RT05)

- HMM was trained on about 100h hours of meeting data.
- NN was trained on 10h subset HMM training.
- Tested on RT05 eval data.
- Based on lattice rescoring of AMI 2005 speech recognizer.

System	WER [%]
PLP SR-HLDA	28.7
PLP SR-HLDA + LCRC-posteriors	26.0

Out of the paper - Comparison of PLPs and LCRC posterior features

- HMM was trained on about 100h hours of meeting data.
- NN was trained on 30h subset HMM training.
- Tested on RT05 eval data.
- Based on lattice rescoring of AMI 2005 speech recognizer.

System	PLP SR-HLDA WER [%]	+ LCRC-posteriors WER [%]
Basic HMM	28.7	25.2
SAT	27.6	23.9
SAT MPE	24.5	21.7

Conclusions

- Smoothing of HLDA bring improvement even on systems with hundreds of training data.
- Remove all silence statistics in HLDA estimation is cheep method to increase performance.
- HLDA on a top of posterior feature generation outperform classic PCA.
- The novel structure of LCRC posterior feature extraction bring a significant WER reduction in both CTS and meeting data experiments.