Defining the Controlling Parameter in Constrained Discriminative Linear Transform for Supervised Speaker Adaptation

Presented by Dimitri Kanevsky from IBM T.J. Watson Research Center

May, 2011 at Prague

Co-authors: Danning Jiang, Emmanuel Yashchin, Yong Qin

Problem Statement

– Discriminative techniques have been proven effective in acoustic model training, and also been investigated in speaker adaptation.
– But, discriminative adaptation is less stable compared with the ML-based techniques:
  - More sensitive to errors presented in the hypothesis.
  - Even for supervised adaptation inappropriate settings of the controlling parameter can cause failures.
– In the presentation, we will:
  - Investigate how the controlling parameter affects the performance of CDLT,
  - And propose a log-linear method to define the parameter.

Recap of CDLT (1)

- Similar to CMLLR, CDLT is also applied at the feature end:

$$\hat{o}(t) = Ao(t) + b = W\zeta(t)$$

- Given a discriminative objective function, the sufficient statistics required to estimate the i-th row of the transform are as follows:

$$\beta = \sum_{j,m}\sum_{t}(\gamma_{jm}^{num}(t) - \gamma_{jm}^{den}(t)) + D_{jm}$$

$$\mathrm{G}^{(\mathrm{i})} = \sum_{j,m}\frac{1}{\sigma_{jm}^{(i)^2}}(\sum_{t}\gamma_{jm}^{num}(t)\zeta(t)\zeta(t)^T - \sum_{t}\gamma_{jm}^{den}(t)\zeta(t)\zeta(t)^T + D_{jm}\begin{bmatrix} 1 & \tilde{\mu}_{jm}^T \\ \tilde{\mu}_{jm} & \tilde{\Sigma}_{jm} + \tilde{\mu}_{jm}\tilde{\mu}_{jm}^T \end{bmatrix}$$

$$k^{(\mathrm{i})} = \sum_{j,m}\frac{\mu_{jm}^{(i)}}{\sigma_{jm}^{(i)^2}}(\sum_{t}\gamma_{jm}^{num}(t)\zeta(t) - \sum_{t}\gamma_{jm}^{\mathrm{den}}(t)\zeta(t) + D_{jm}\begin{bmatrix} 1 \\ \tilde{\mu}_{jm} \end{bmatrix}$$

Recap of CDLT (2)

- The usual way to set the smoothing factor:

$$D_{jm} = E\sum_{t}\gamma_{jm}^{den}(t)$$

- E is the learning rate and usually set as a constant inside [1.0,2.0], which is the same with that used in discriminative training.
- However, the E used in discriminative adaptation can be different:
  - The learning rate should be aggressive enough for efficient adaptation.
  - But can't be so aggressive that the EBW optimization is failed.
  - Usually there is no "development" data in adaptation and it is hard to decide what learning rate should be used and when the iteration should be stopped.

Effects of the controlling parameter

- We empirically studied effects of the controlling parameter in supervised CDLT adaptation:

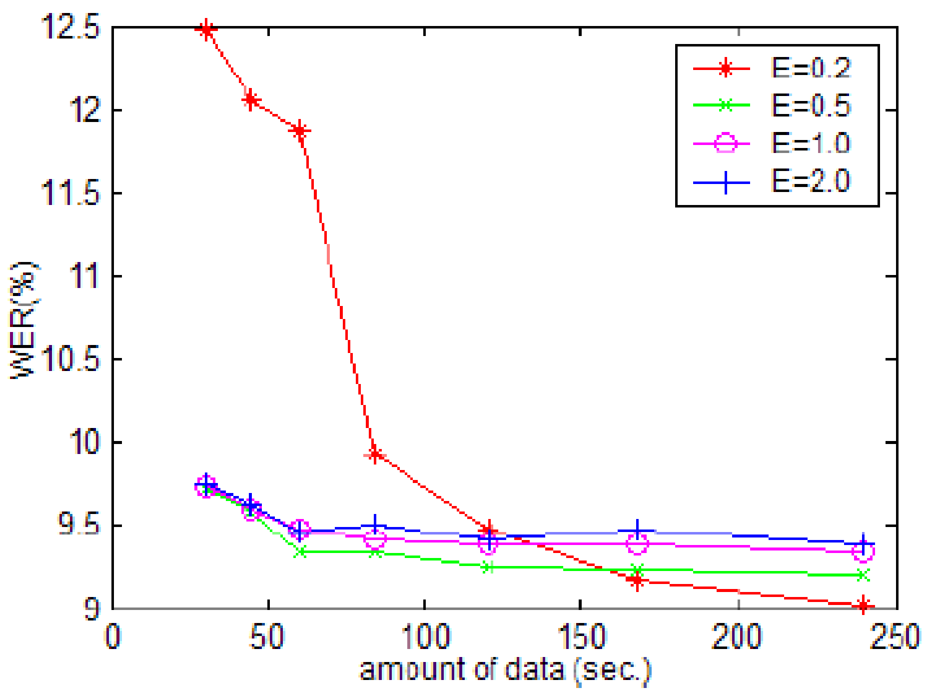

Figure 1. Comparison of CDLT performance with E=0.2, 0.5, 1.0, 2.0 when the amount of adaptation data varies (on test set 1)

- 26 speakers in the data set. Up-to-4 minutes' enrollment data per speaker, and 7-20 minutes' test data per speaker.

- Boosted MMI criterion is used in the CDLT estimation.

- Apparently, the optimal E highly depends on the amount of adaptation data.

The log-linear dependence

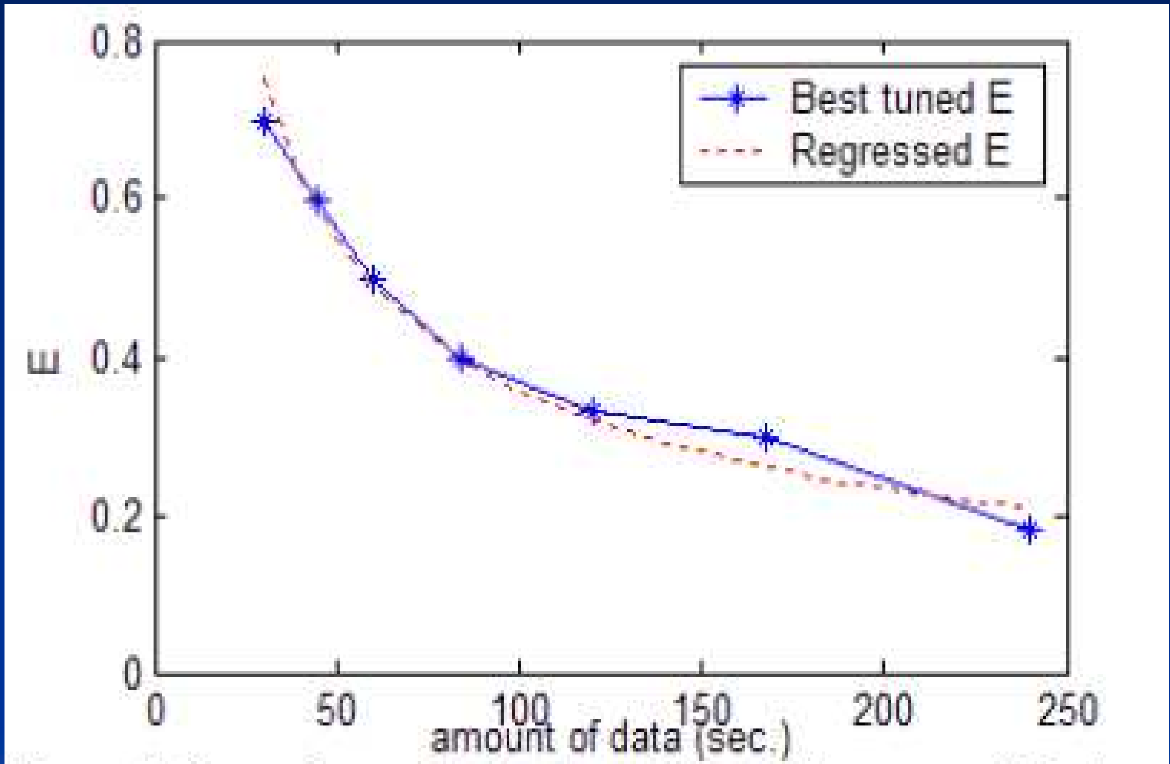

Figure 2. Comparison of the best tuned E and the regressed E when the amount of adaptation data varies (on test set 1).

- The dependence of the best E on the amount of adaptation data tends to be log-linear (as figure 2 shows).
- The correlation coefficient of ln(E) and ln(n) is -0.98.
- The empirical formula obtained via linear regression:

$$\ln(E) = 1.802 - 0.618 * \ln(n)$$

Recognition experiments (1)

- Test data
  - Test set 1: 26 speakers (the data set used to derive the log-linear formula).
  - Test set 2: 21 speakers.
  - In both sets, up-to-4 minutes' adaptation data and 7-20 minutes' test data for each speaker.
- The base model contains 5k tied-states and 200k Gaussian components, trained on 2000 hours of data.
- SAT training was first applied on the LDA features, where the speaker-specific transforms were estimated via CMLLR.
- fMPE and MPE training were finally performed based on the SAT model.

Recognition experiments (2)

Table 1. WERs of CDLT with different E setting methods and CMLLR on test set 1.

	CMLLR	CDLT		
		E=0.5	Tuned E	Predicted E
30 sec.	9.74	9.71	9.67	9.70
45 sec.	9.70	9.58	9.57	9.57
1.0 min.	9.52	9.34	9.34	9.34
1.4 min.	9.51	9.34	9.29	9.29
2.0 min.	9.50	9.25	9.16	9.16
2.8 min.	9.44	9.24	9.12	9.15
4.0 min.	9.45	9.21	9.02	9.03

Table 2. WERs of CDLT with different E setting methods and CMLLR on test set 2.

	CMLLR	CDLT		
		E=0.5	Tuned E	Predicted E
30 sec.	7.20	7.26	7.14	7.14
45 sec.	7.21	7.18	7.14	7.15
1.0 min.	7.13	7.12	7.08	7.12
1.4 min.	7.14	7.01	7.00	7.03
2.0 min.	7.11	7.00	6.95	6.96
2.8 min.	7.04	6.94	6.87	6.87
4.0 min.	7.03	6.97	6.82	6.82

- CMLLR and CDLT with three learning rates are compared:
  - Set as a fixed E (E=0.5).
  - Defined by the log-linear formula.
  - Manually tuned.
- For both sets, CDLT with the E defined by the log-linear formula clearly outperformed CMLLR and the CDLT baseline (E=0.5).

More discussions on the log-linear dependence (1)

- Does it exist for general classes of similar models?
  - It is possible.
  - For example, we can prove it in Ridge regression. Let the regression model be:

    $$Y = X\beta + \varepsilon$$

    It has been proven that to achieve the best predictive ability:

    $$\text{Minimize} \quad (1/n)\| Y\text{-}Xb \|^2 + \lambda \| b \|^2$$

    where λ is a controlling parameter similar with E in CDLT. It is proven in the paper that $d = \lambda * n$ tends to be a constant as the sample size increases, implying a relationship of the following formula:

    $$\ln(\lambda) = c0 - \ln(n) + o(\ln(n))$$

Conclusions

- An empirical study that investigates impacts of the EBW controlling parameter on the adaptation performance of CDLT
- A log-linear relationship exists between the optimal setting of the controlling parameter E and the amount of adaptation data
- With E set based on the log-linear relationship, CDLT performance was better than the CDLT baseline
- Proved that the log-linear relationship does exist for Ridge regression
- we can expect that the log-linear relationship holds more generally in multiple settings, since regularized linear regression is the backbone of many learning problems