香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

THE DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

An Investigation of Adaptation Techniques for Building Acoustic Models for Hearing-impaired Children in a CAPT Application

Yingke Zhu Brian Mak
{yzhuav, mak} @cse.ust.hk

ISCSLP 2016@Tianjin

Outline

- Computer assisted pronunciation training system
- Task description
- Adaptation techniques
  - KL divergence regularization
  - Linear input network
  - Learning hidden-unit contributions (LHUC)
- Evaluation in the real system

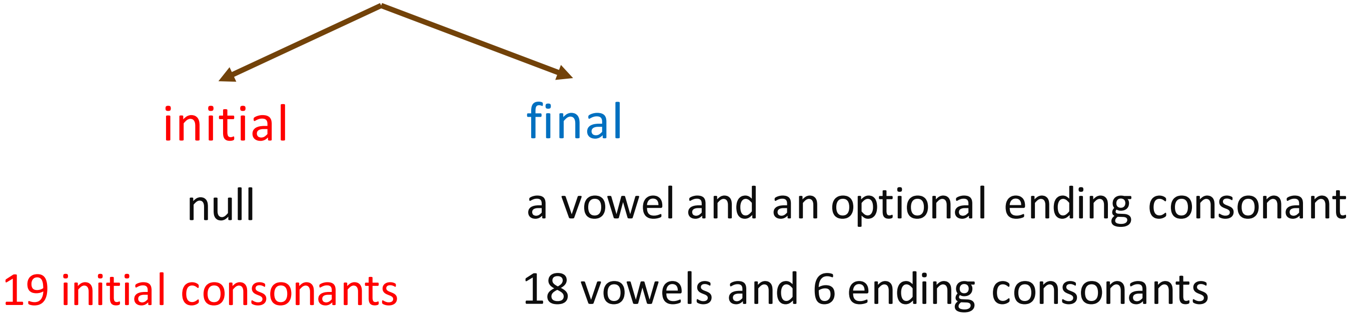
▪ CAPT system

- Android-based computer-assisted pronunciation training application developed for the local hearing-impaired (HI) children.

- Contains listening and speaking exercises of around 400 Cantonese words.

- Cantonese

Each character is a syllable.



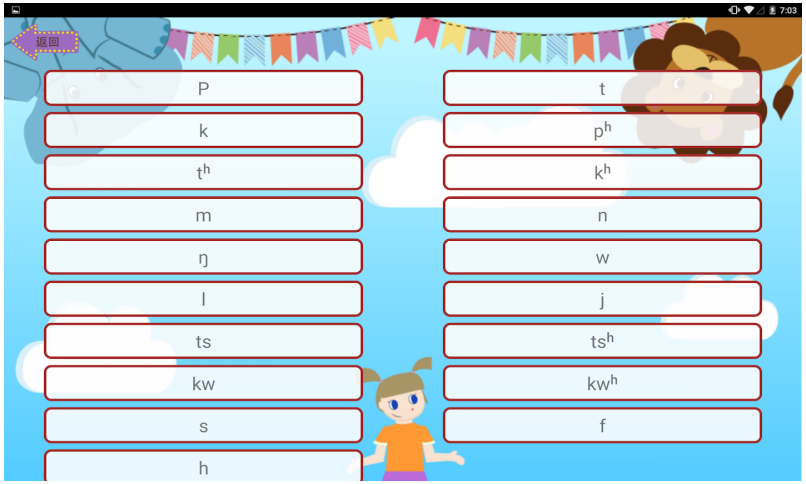
initial	final
null	a vowel and an optional ending consonant
19 initial consonants	18 vowels and 6 ending consonants

- CAPT system
  - Exercise

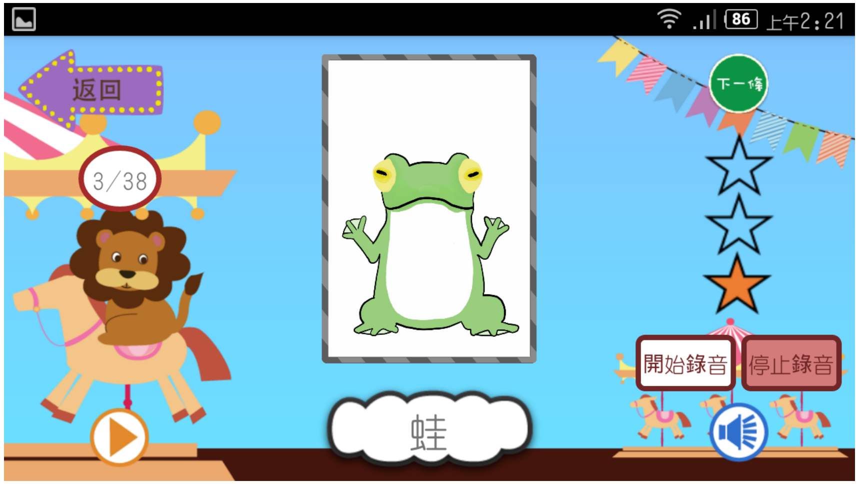
Tell the difference between two very similar words that differ only in their initial consonants.

/tɒu/ → [kɒu]

豆(bean) 狗(dog)



Exercise for 19 initial consonants



Assessment

Phoneme verification problem

Task Description

- Aim

  Score the pronunciation of the initial phone in Cantonese words.

- Acoustic model

  Modeling Cantonese monophones

- Performance metrics

  PER: overall phone error rate

  ICER: initial consonant error rate

- Acoustic Model

Target user

Hearing impaired(HI) children in Hong Kong

- Aged between 6 to 12

Problem

Lack of data

- 1 hour of Cantonese speech from 36 HI children

Strategy

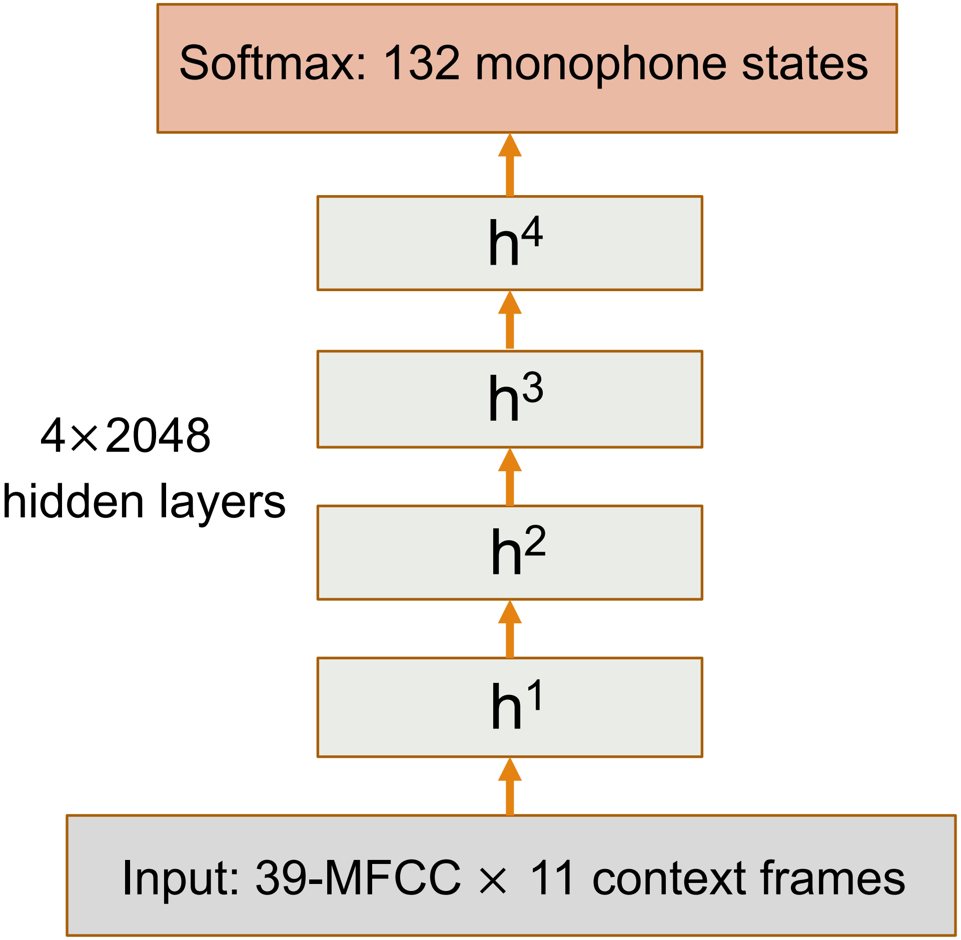
Training → NH adults acoustic model

- Sufficient normal hearing adults speech data

Group Adaptation → HI children acoustic model

- HI children speech data

Acoustic Model



Softmax: 132 monophone states

h^4

h^3

h^2

h^1

4×2048
hidden layers

Input: 39-MFCC × 11 context frames

Training: NH adult model

- 35 hours speech data from 166 normal hearing adults

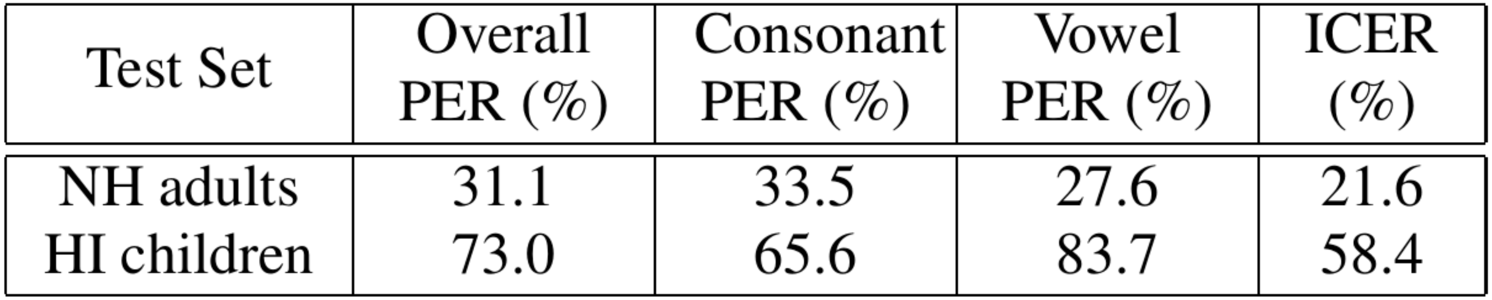
Adaptation: HI children model

- 1 hours speech data from 36 hearing impaired children

Data set	# Speakers	Amount
Adaptation	18	0.51 h
Dev	9	0.22 h
Test	9	0.27 h

▪ NH Adult Acoustic Model

- Results on two test sets



Test Set	Overall PER (%)	Consonant PER (%)	Vowel PER (%)	ICER (%)
NH adults	31.1	33.5	27.6	21.6
HI children	73.0	65.6	83.7	58.4

- The performance of NH adult acoustic model on HI children test set has a significant drop.

- There's a mismatch between two speech corpus.

Adults – Children

Normal hearing – Hearing impaired

Adaptation techniques

- KL divergence regularization
- Linear input network
- Learning hidden-unit contributions

KL divergence regularization

- DNN training – optimization criterion

$$\overline{D} = \frac{1}{N}\sum_{t=1}^{N} D(\mathbf{x}_t) = \frac{1}{N}\sum_{t=1}^{N}\sum_{y=1}^{S} \underbrace{\widetilde{p}(y|\mathbf{x}_t)}_{= \begin{cases} 1 & \text{if } y = s_t \\ 0 & \text{otherwise} \end{cases}} \log p(y|\mathbf{x}_t)$$

- KLD adaptation – regularized optimization criterion

$$\hat{D} = (1-\rho)\overline{D} + \rho \cdot \frac{1}{N}\sum_{t=1}^{N}\sum_{y=1}^{S} p^{SI}(y|\mathbf{x}_t) \log p(y|\mathbf{x}_t)$$

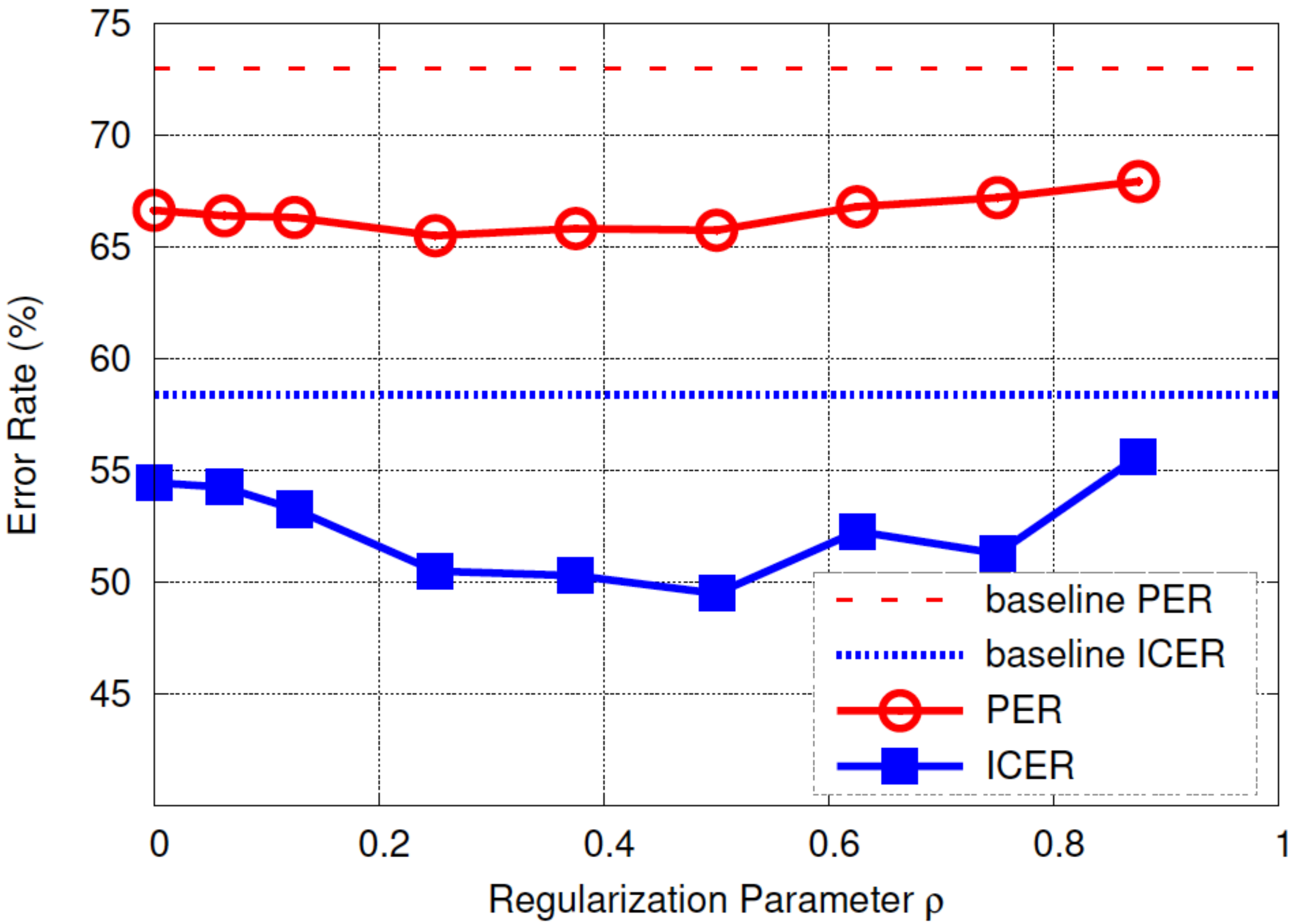
- Implementation

$$\hat{D} = \frac{1}{N}\sum_{t=1}^{N}\sum_{y=1}^{S} \underbrace{\hat{p}(y|\mathbf{x}_t)}_{\Big\downarrow} \log p(y|\mathbf{x}_t)$$

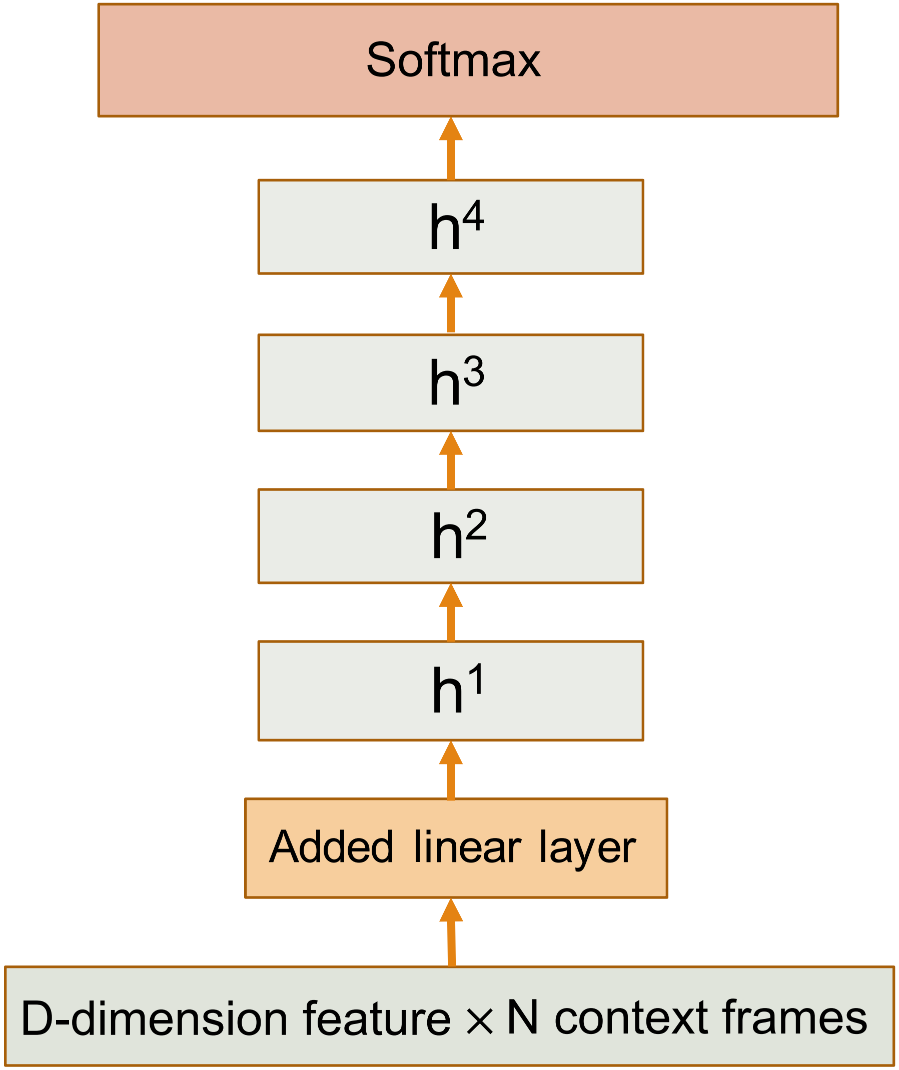
$$(1-\rho)\cdot\widetilde{p}(y|\mathbf{x}_t) + \rho \cdot p^{SI}(y|\mathbf{x}_t)$$

Conventional BP algorithm

- KL divergence regularization



▪ Linear input network



Softmax

h^4

h^3

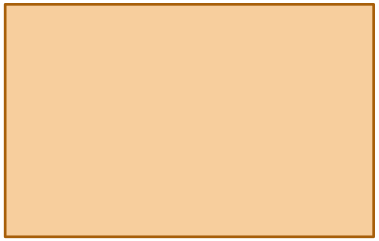
h^2

h^1

Added linear layer

D-dimension feature × N context frames

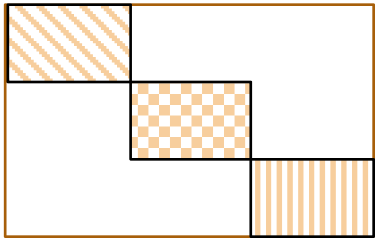
- LIN



intra-frame and inter-frame relations

#Parameters: ND× (ND+1)

- LIN-Nblock

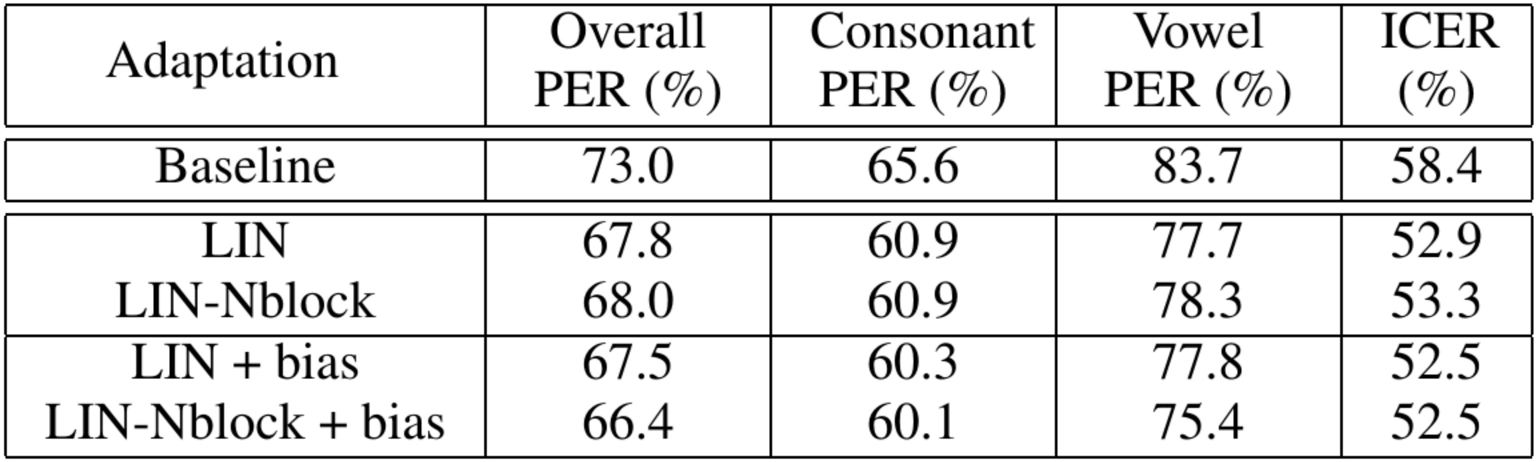


intra-frame relation

#Parameters: ND× (D+1)

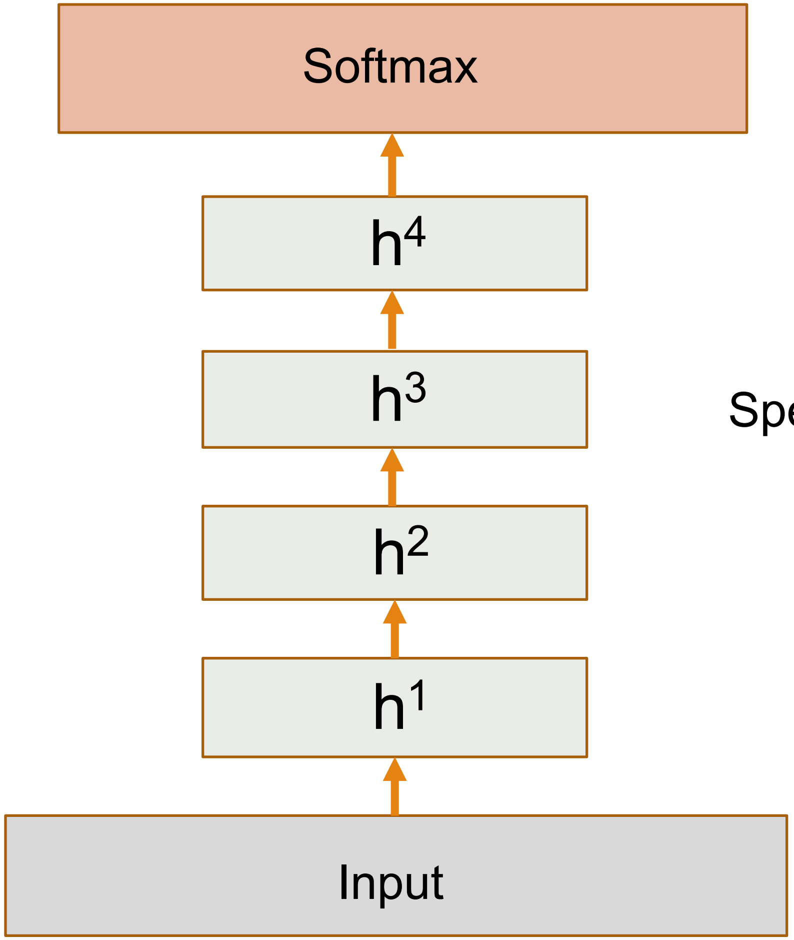
Linear input network

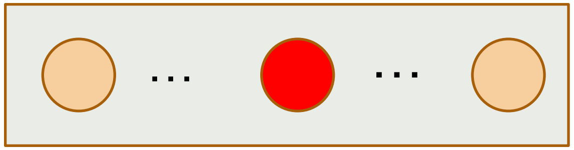
- Results



Adaptation	Overall PER (%)	Consonant PER (%)	Vowel PER (%)	ICER (%)
Baseline	73.0	65.6	83.7	58.4
LIN	67.8	60.9	77.7	52.9
LIN-Nblock	68.0	60.9	78.3	53.3
LIN + bias	67.5	60.3	77.8	52.5
LIN-Nblock + bias	66.4	60.1	75.4	52.5

- Learning hidden-unit contributions (LHUC)





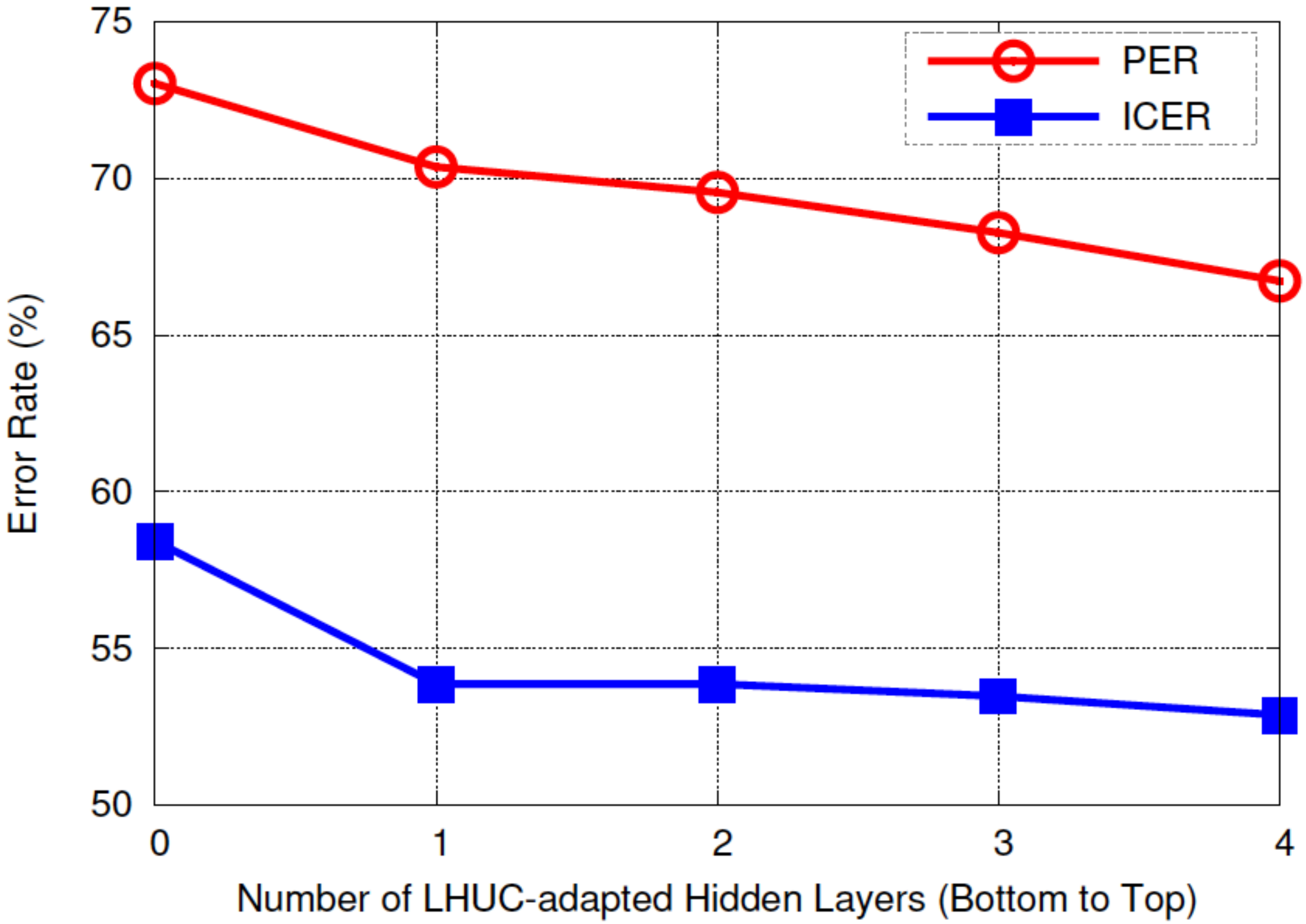
Speaker independent DNN: $h_i^l = \sigma(z_i^l)$

LHUC adaptation: $h_i^l = a_i^l \cdot \sigma(z_i^l)$

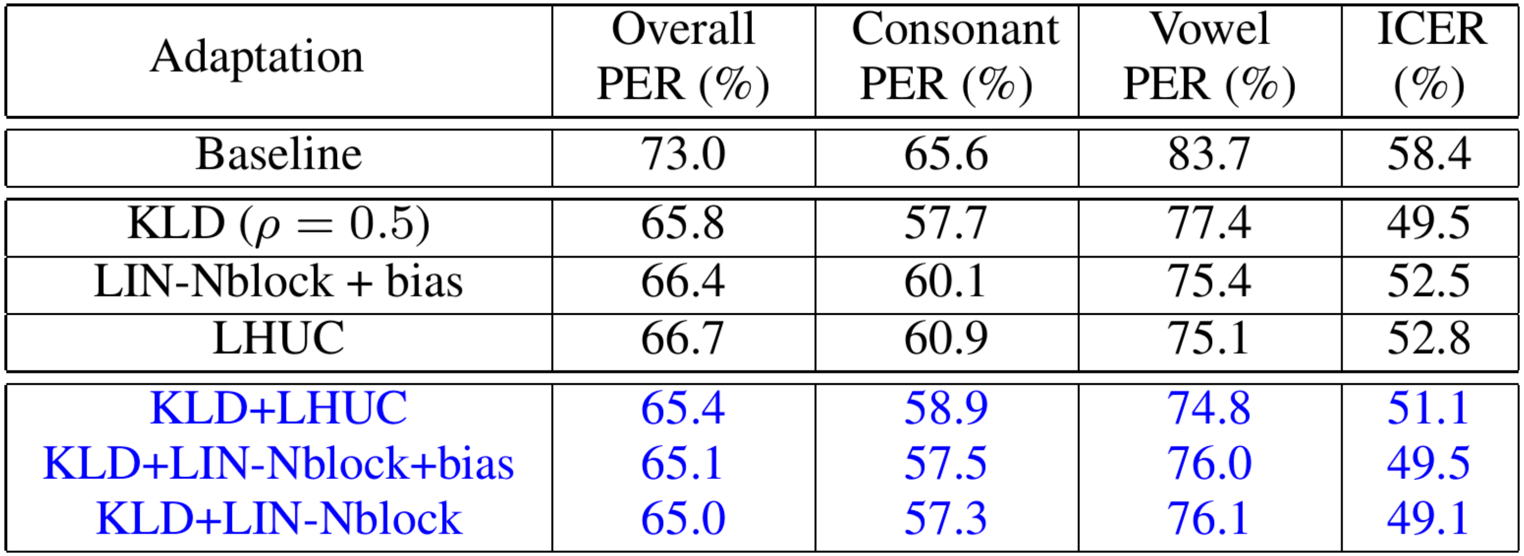
Choice of scaling factor: $a_i^l = \frac{2}{1+e^{-r_i^l}}$

#Parameters: #adapted hidden layers × #hidden nodes

- Learning hidden-unit contributions (LHUC)

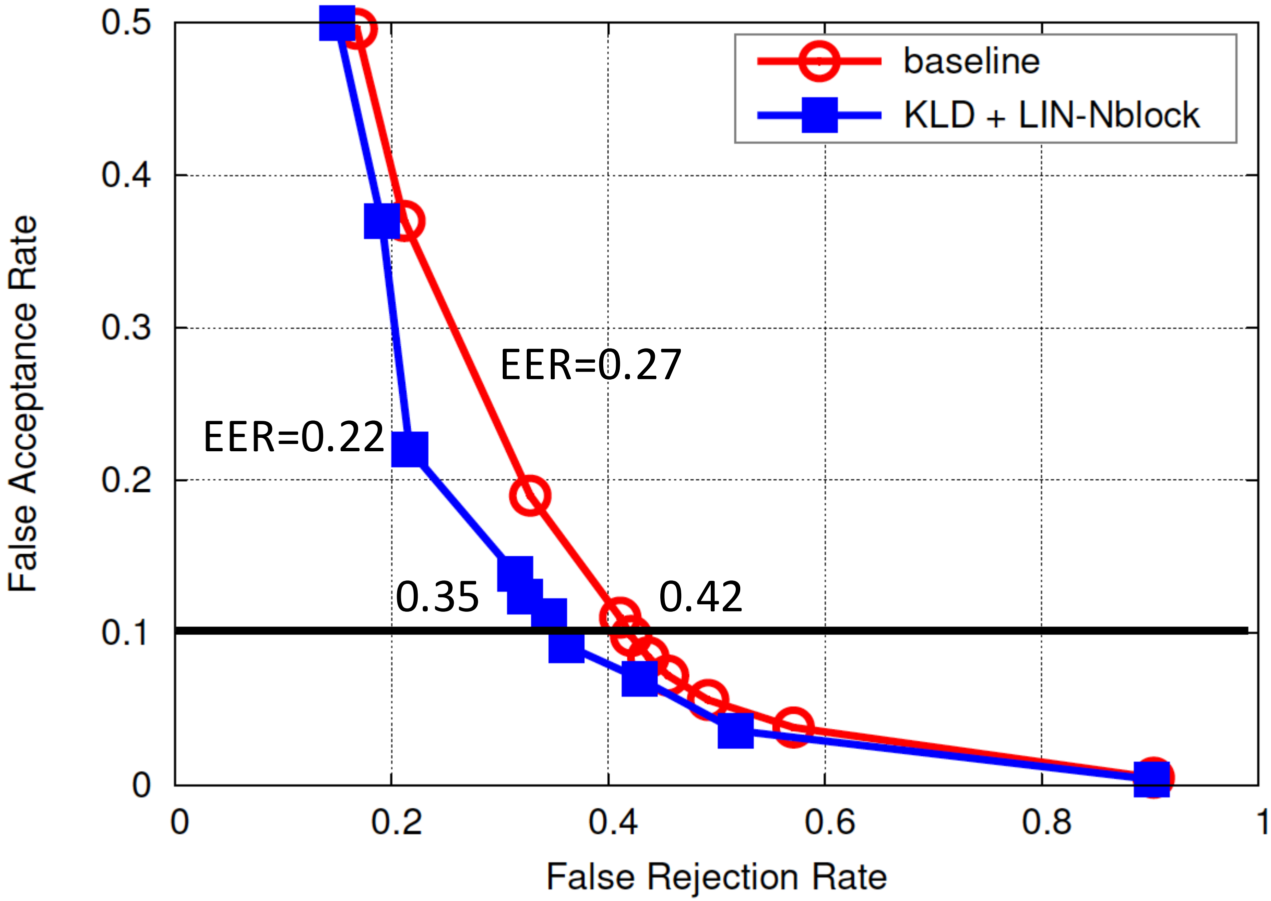


- Adaptation results summarization



Adaptation	Overall PER (%)	Consonant PER (%)	Vowel PER (%)	ICER (%)
Baseline	73.0	65.6	83.7	58.4
KLD ($\rho = 0.5$)	65.8	57.7	77.4	49.5
LIN-Nblock + bias	66.4	60.1	75.4	52.5
LHUC	66.7	60.9	75.1	52.8
KLD+LHUC	65.4	58.9	74.8	51.1
KLD+LIN-Nblock+bias	65.1	57.5	76.0	49.5
KLD+LIN-Nblock	65.0	57.3	76.1	49.1

- Evaluation in the real system
  - Assessment performance is reported in terms of equal error rate (EER).



System operation point

Conclusion

- We investigated various speaker adaptation techniques for group adaptation: adapting an NH adults acoustic model to work force HI children in a mobile CAPT application.
- The major challenges are:
  - the acoustic characteristics of HI children speech are very different from those of NH speakers in the original model
  - the amount of adaptation data is very limited
- We investigated KLD regularization, LHUC, LIN, and their combinations. Among the three methods, if they were applied alone, KLD regularization gave the best performance.
- Further improvement could be achieved from the joint adaptation of KLD and LIN-Nblock, reducing PER and ICER by a relative 11% and 16% respectively.

Q & A