TOSHIBA
Leading Innovation >>>

Statistical Parametric Speech Synthesis Based on Speaker & Language Factorization

Heiga ZEN

Toshiba Research Euorpe Ltd.
Cambridge Research Lab.

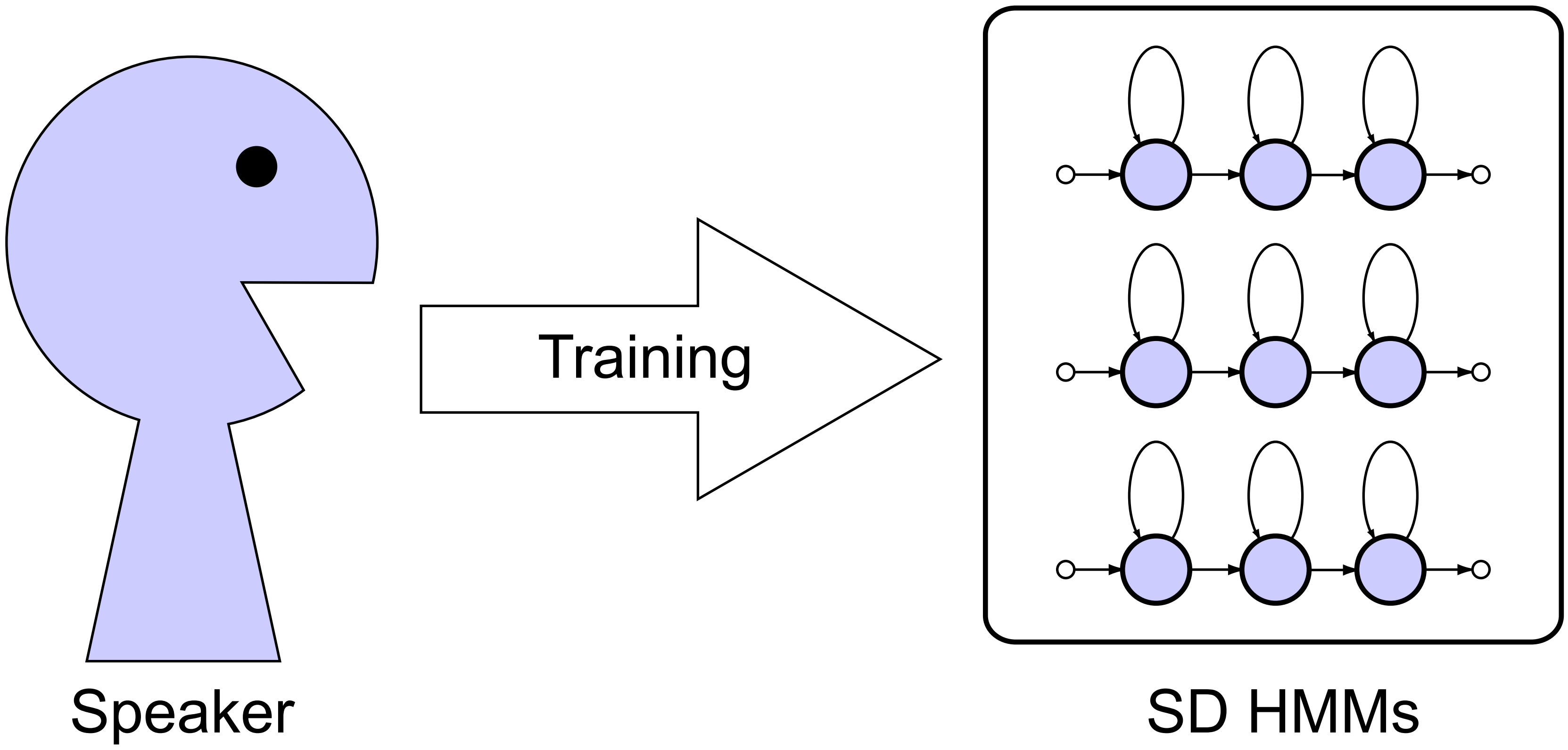
Speech Synthesis Seminar Series @ CUED, Cambridge, UK
June 21st, 2011

Background (1)

Use of inhomogeneous data for training HMMs

- Speech data from single source (e.g., speaker)

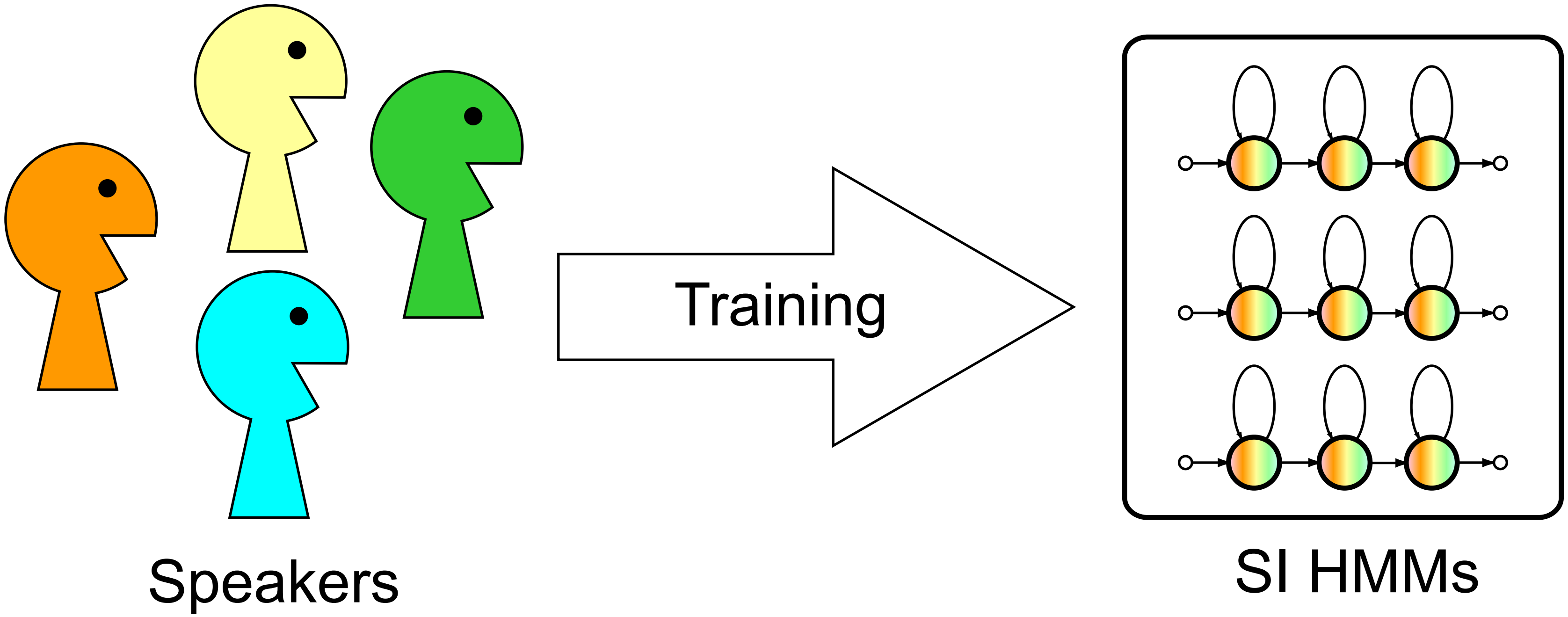
* Amount of available data is limited



Background (1)

Use of inhomogeneous data for training HMMs

- **Speech data from single source (e.g., speaker)**
  * Amount of available data is limited
- **Multi-style learning**
  * Mix speech data from multiple sources

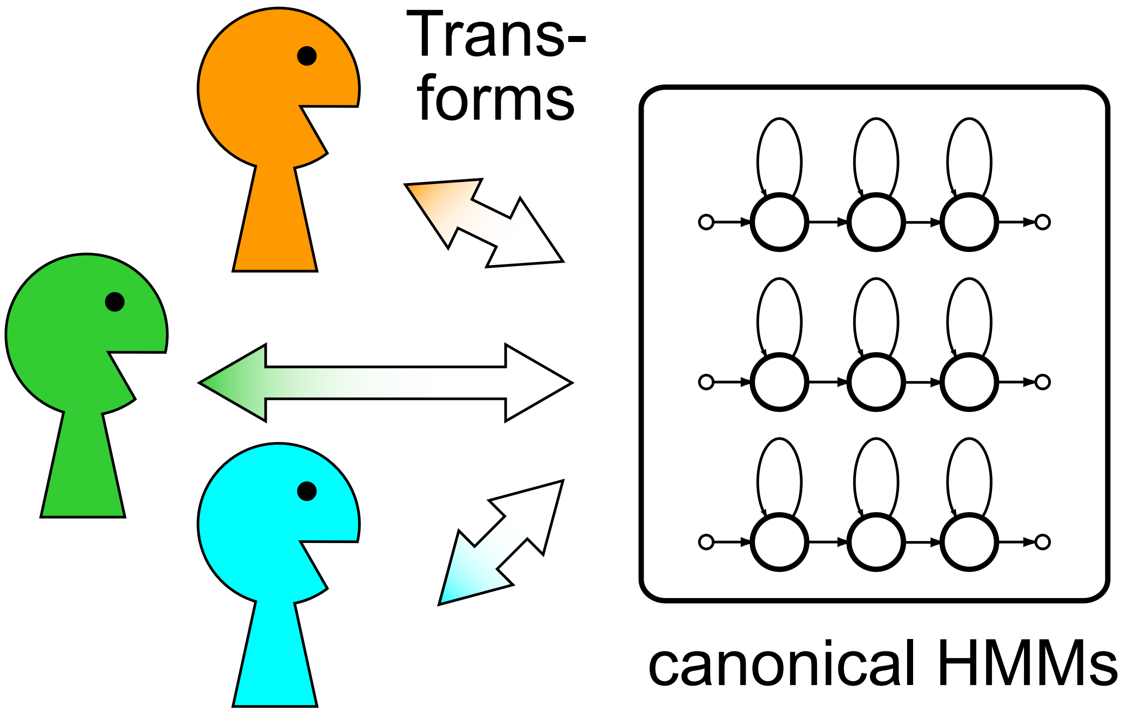


Background (2)

Use of inhomogeneous data for training HMMs

- Adaptive training [Anastasakos;'96]

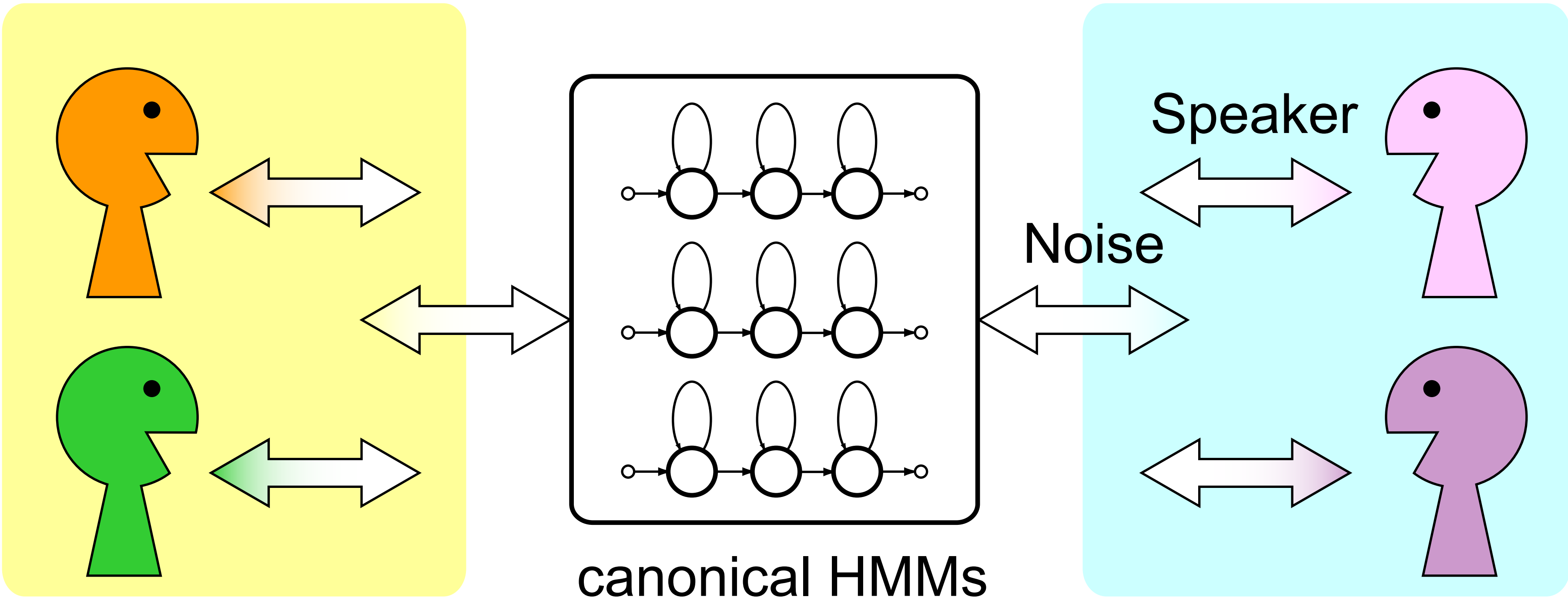
* One transform for each homogeneous block
* Canonical model set is estimated given transforms



Use of inhomogeneous data for training HMMs

- Acoustic factorisation [Gales;'01]

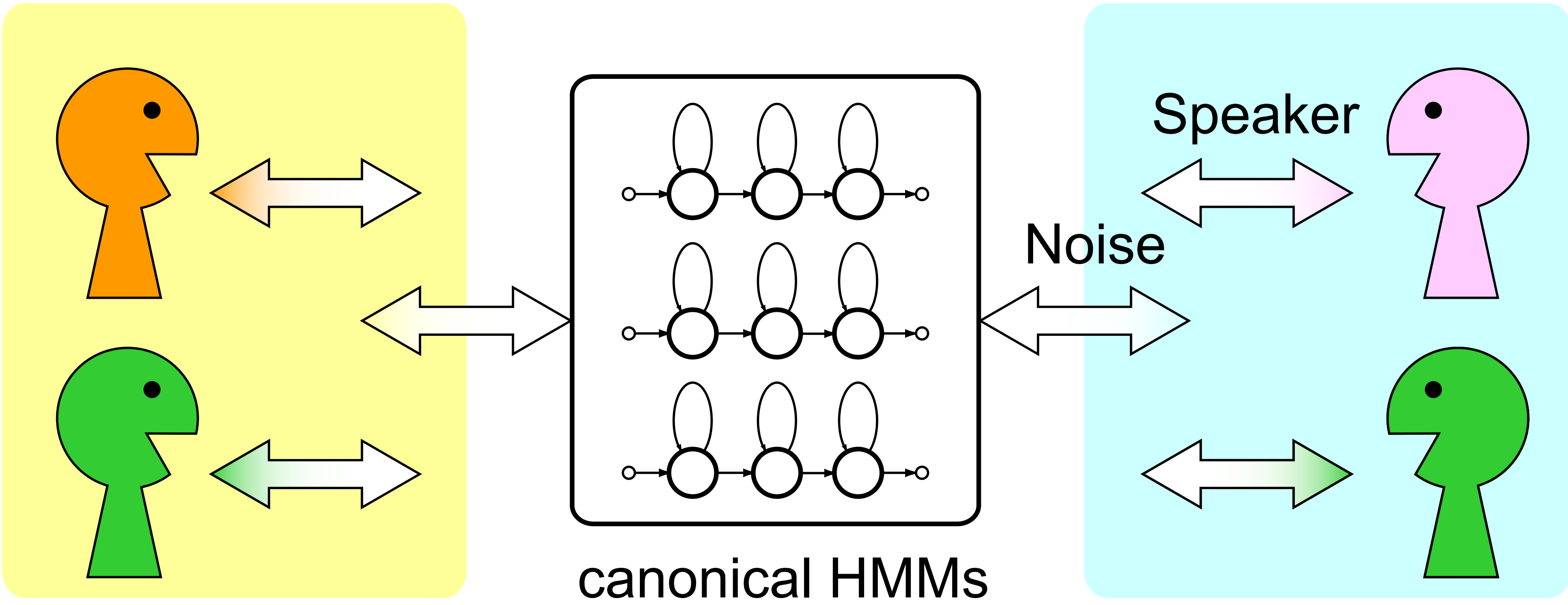
* Multiple factors (e.g., speaker & noise)
* One transform for each factor
* Alter one transform while fixing the others



Background (3)

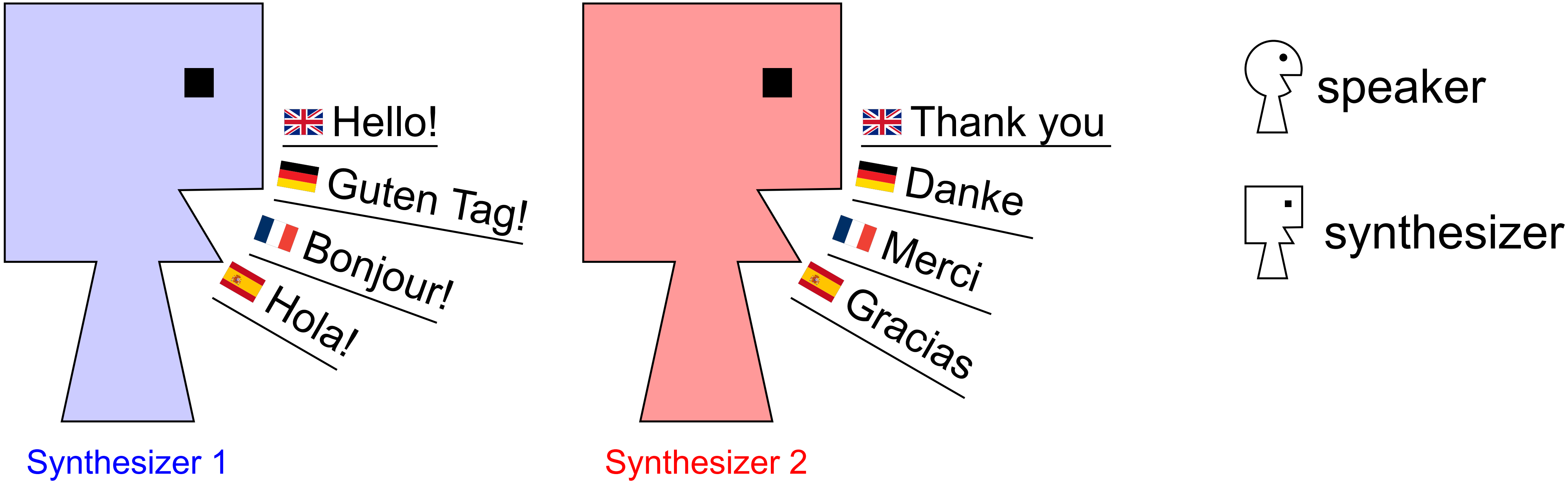
Use of inhomogeneous data for training HMMs

- **Acoustic factorisation [Gales;'01]**
  * Multiple factors (e.g., speaker & noise)
  * One transform for each factor
  * Alter one transform while fixing the others



Polyglot Speech Synthesis

Synthesize multiple languages with common voice

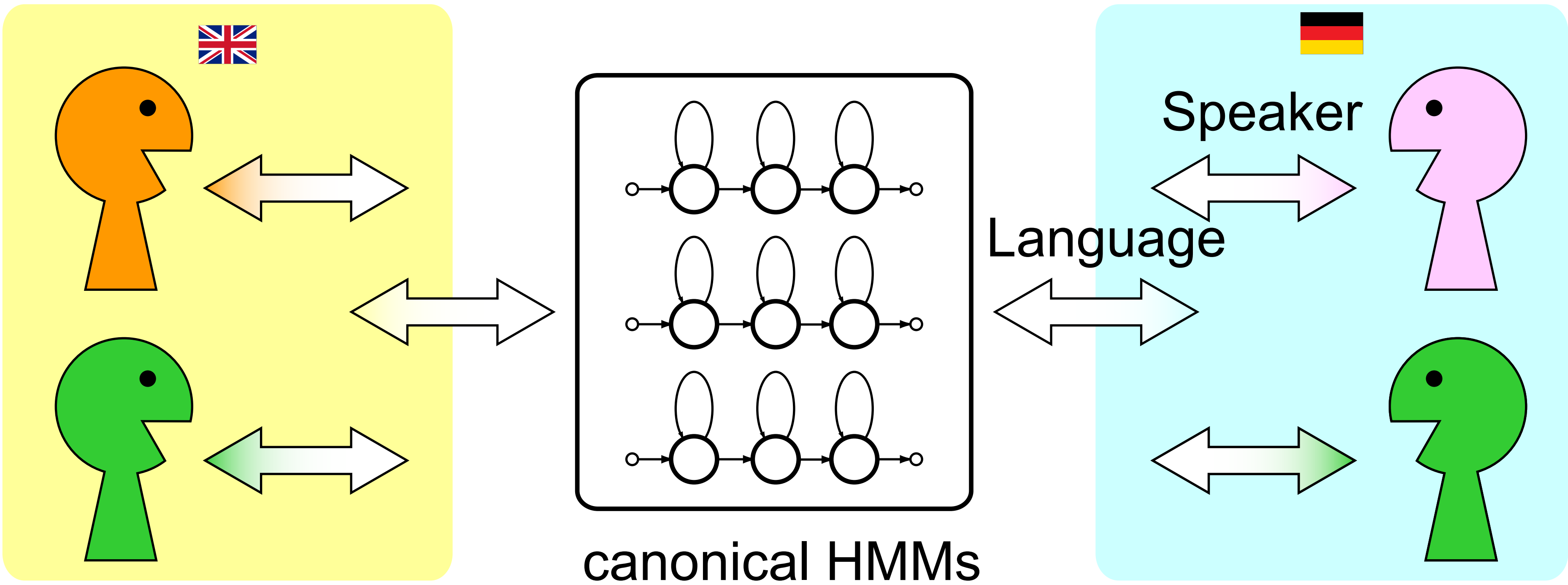


Applications

* Synthesize mix-lingual texts
* Speech-to-speech translators
* More efficient development of TTS for multiple languages

Polyglot Synthesis as Acoustic Factorization

* Two factors (speaker & lang.), one transform for each factor
* Alter language transform with the same speaker transform
  $\Rightarrow$ Polyglot synthesis can be achieved
* Increase amount of data by having multiple languages

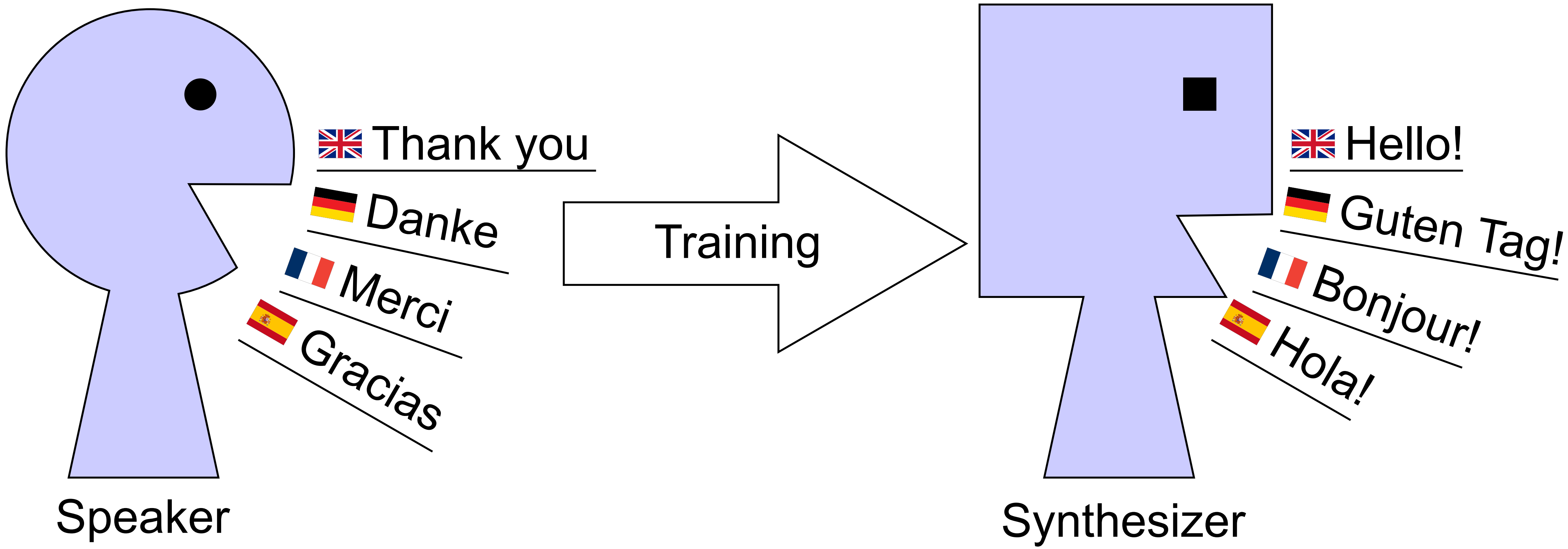


Outline

- Background
- Conventional approaches
    * Polyglot speaker
    * Mixing mono-lingual corpora
    * Cross-lingual speaker adaptation
- Speaker & language factorization (SLF)
    * Concept
    * Details
- Experiments
- Conclusions

Conventional Approaches (1)

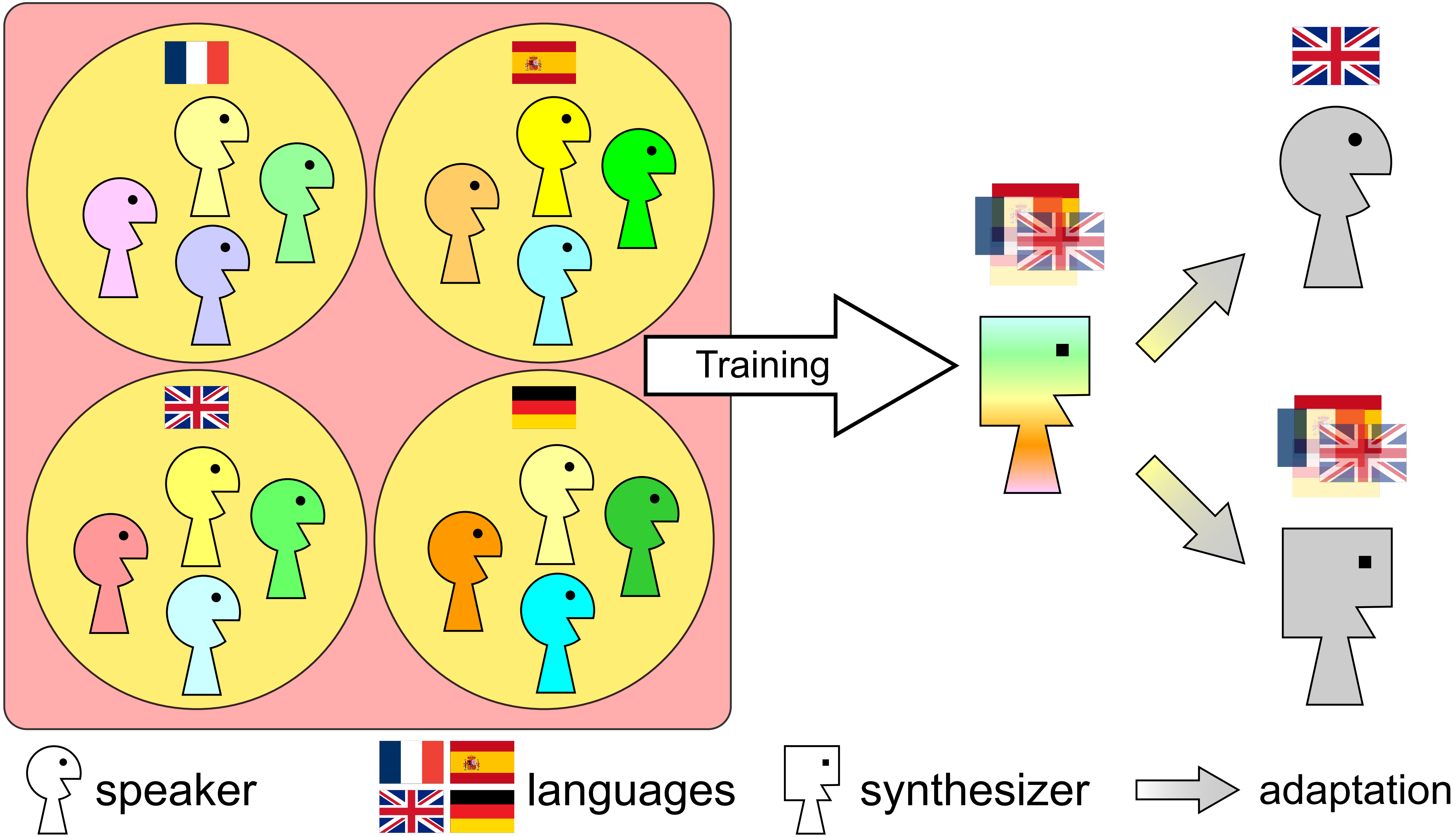
Polyglot speaker [Traber;'99]



Finding good polyglot speakers is very difficult
→ Hardly expandable

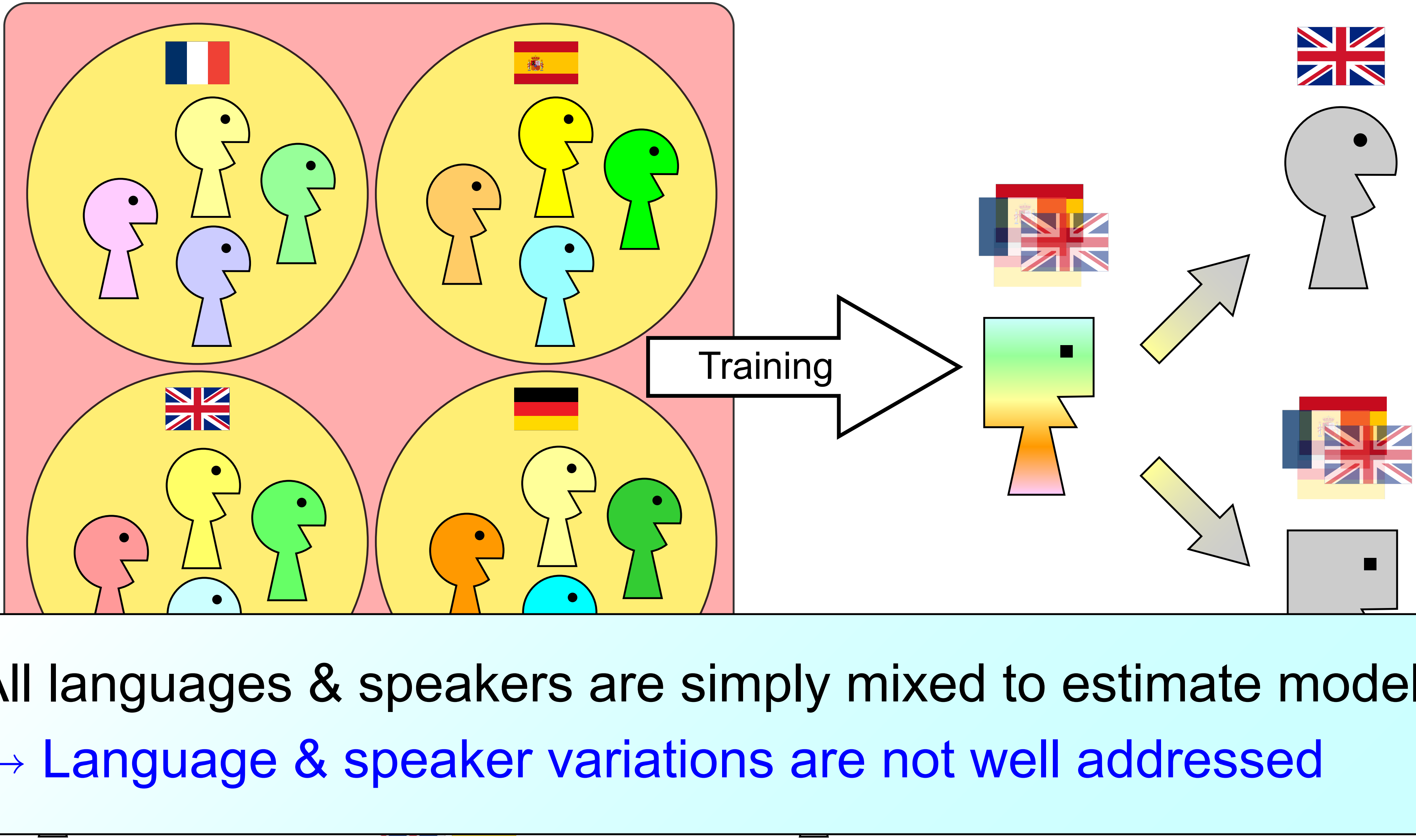
Conventional Approaches (2)

Mix mono-lingual corpus [Latorre;'06, Black;'06]



Conventional Approaches (2)

Mix mono-lingual corpus [Latorre;'06, Black;'06]

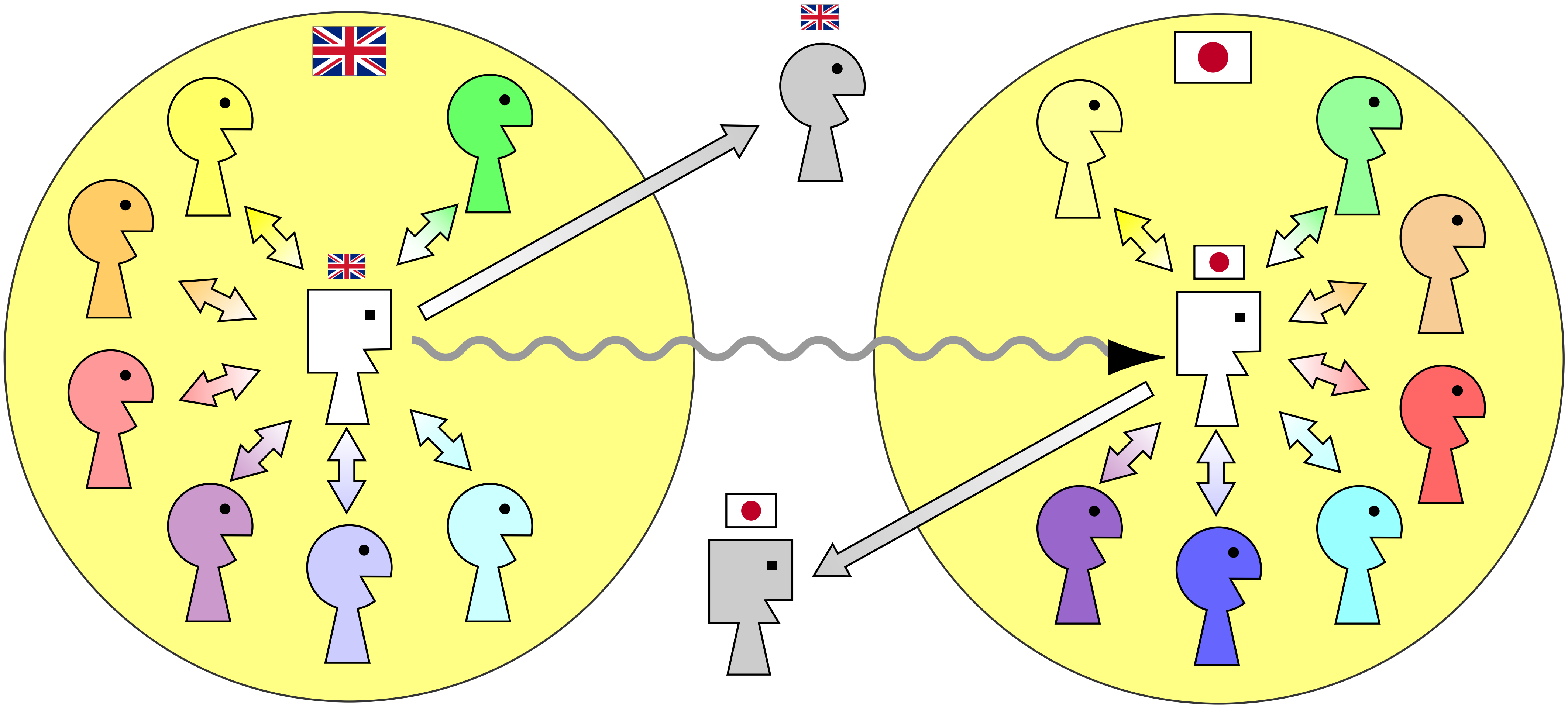


All languages & speakers are simply mixed to estimate model

→ Language & speaker variations are not well addressed

Conventional Approaches (3)

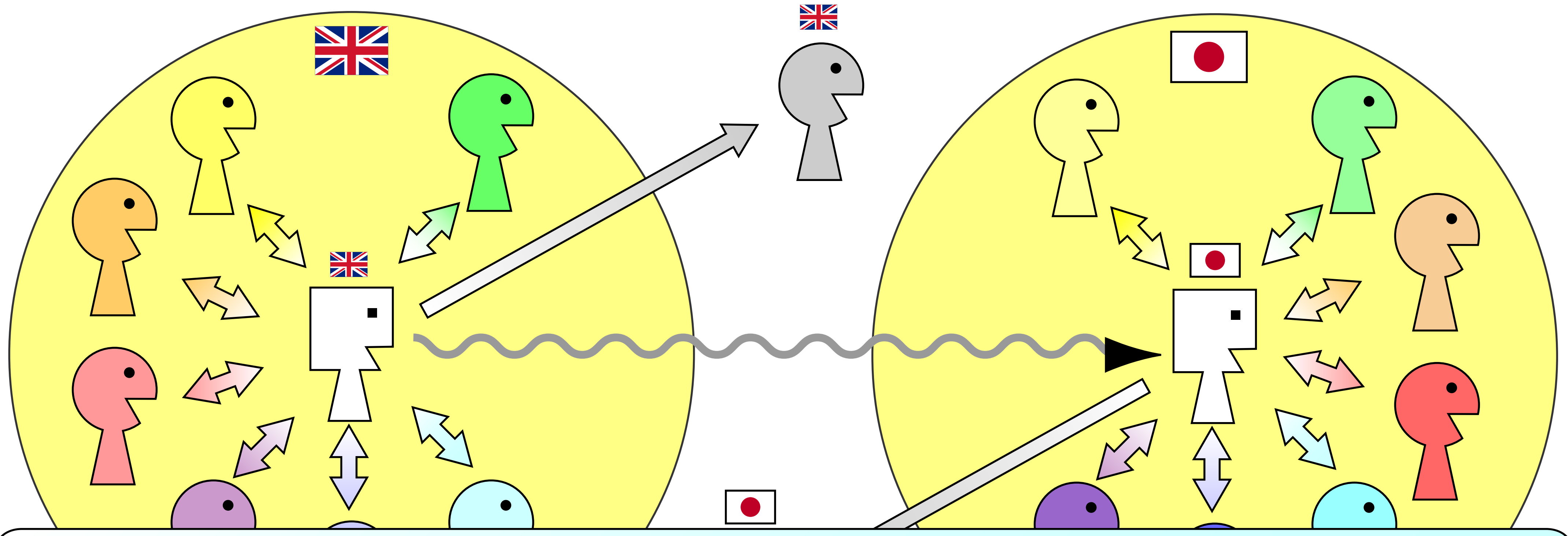
Cross-language speaker adaptation [Chen;'09, Wu;'09]



adaptive training mapping adaptation

Conventional Approaches (3)

Cross-language speaker adaptation [Chen;'09, Wu;'09]

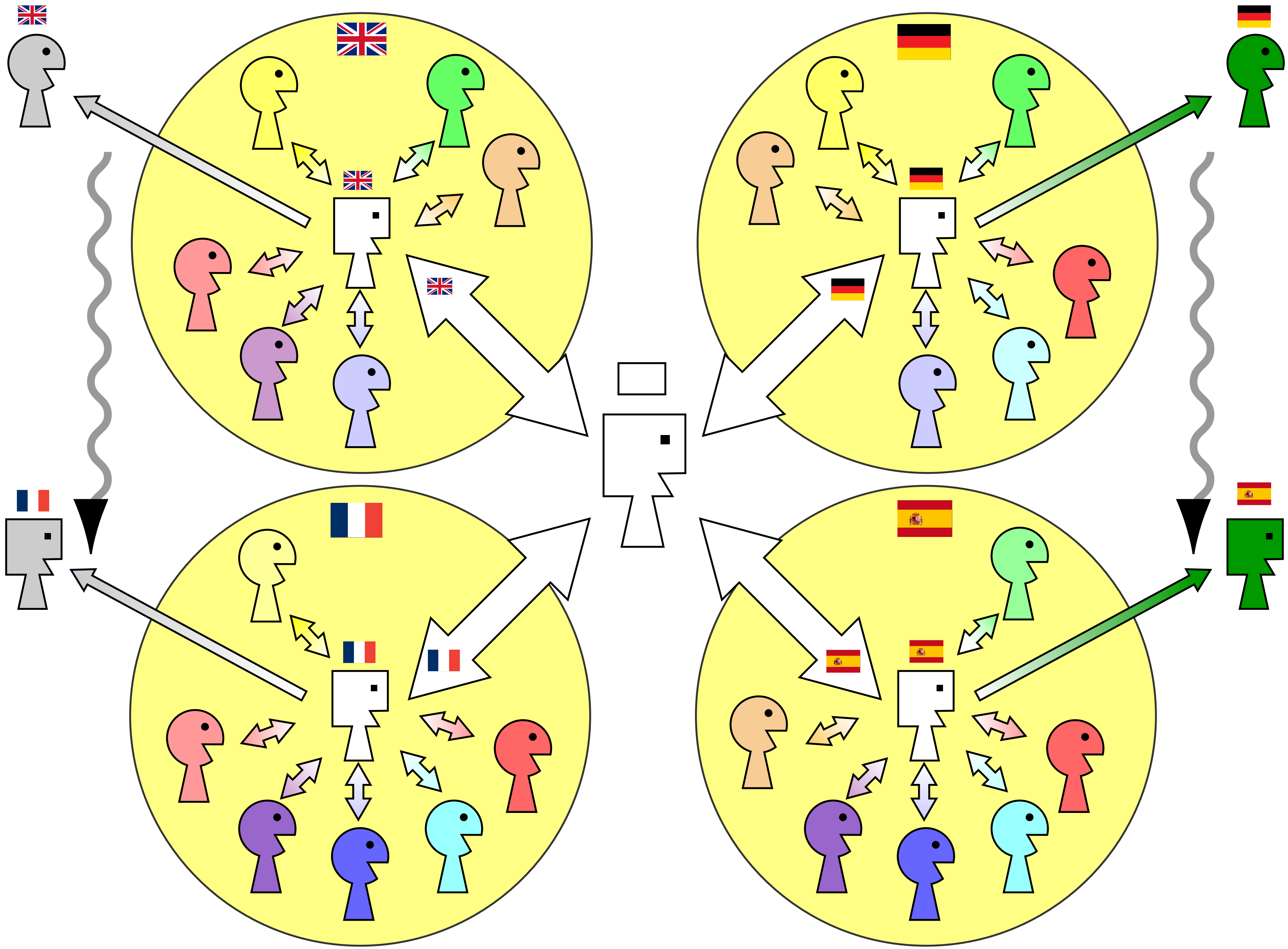


Language-dependent SAT models are estimated independently

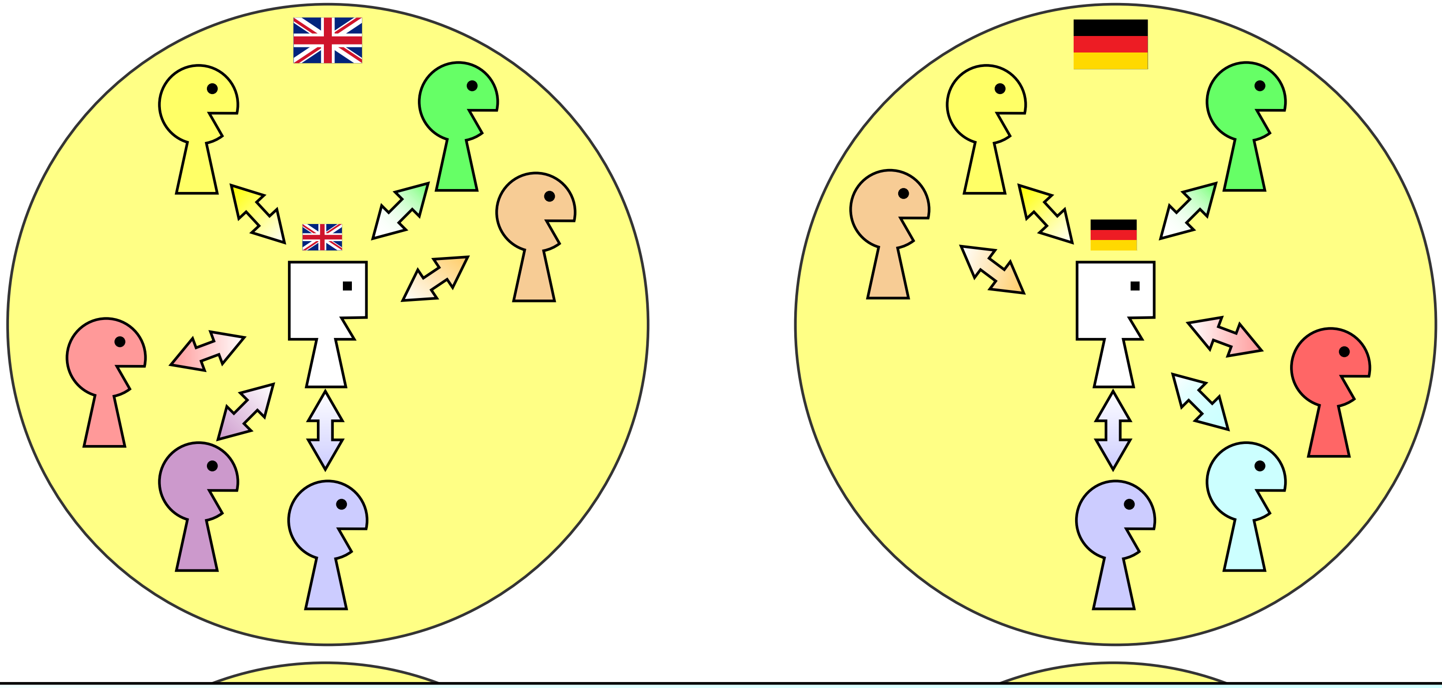
→ Mismatch between language-dependent SAT models

→ Degrade adaptation & synthesis [Liang;'10]

Speaker & Language Factorization (SLF)



Speaker & Language Factorization (SLF)

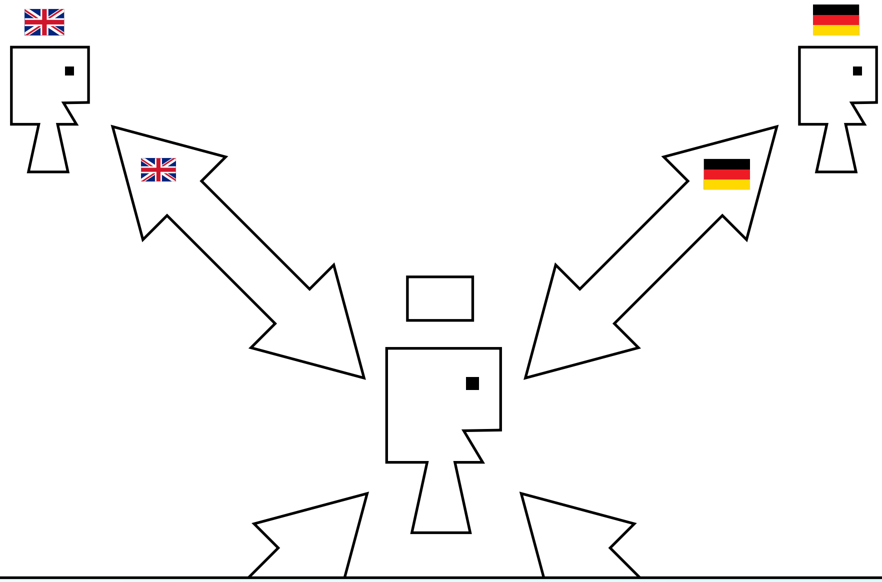


Speaker transform

- Speaker-specific characteristics

* Vocal tract length & shape, F0 height & range, voicing
* Speaking rate
* Speaker-specific speaking styles

Speaker & Language Factorization (SLF)

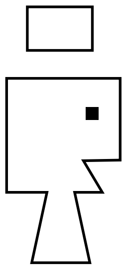


Language transform

- Language-specific characteristics

* Language-dependent parts of syntactic, morphological, intonational, phonetic, & phonological factors

Speaker & Language Factorization (SLF)

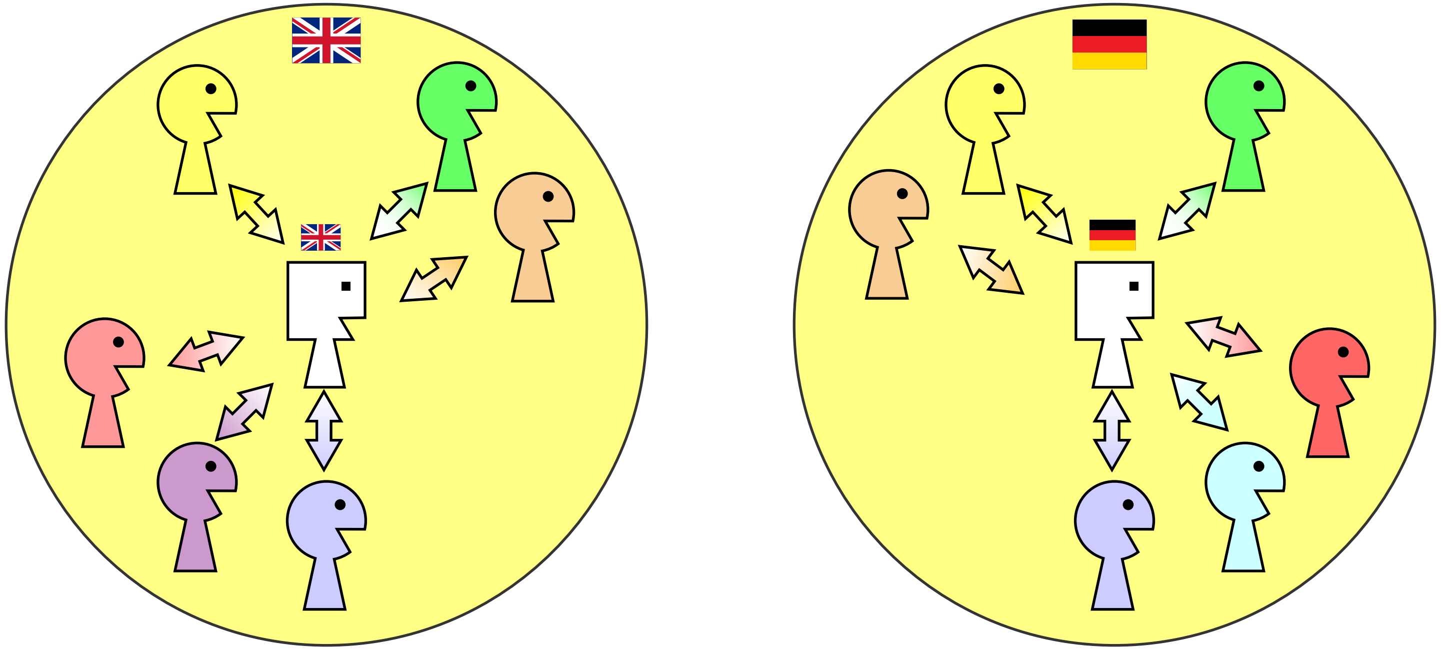


Canonical model

- Common characterisics across languages/speakers

* Cross-language parts of syntactic, morphological, intonational, phonetic, & phonological factors

Speaker & Language Factorization (SLF)



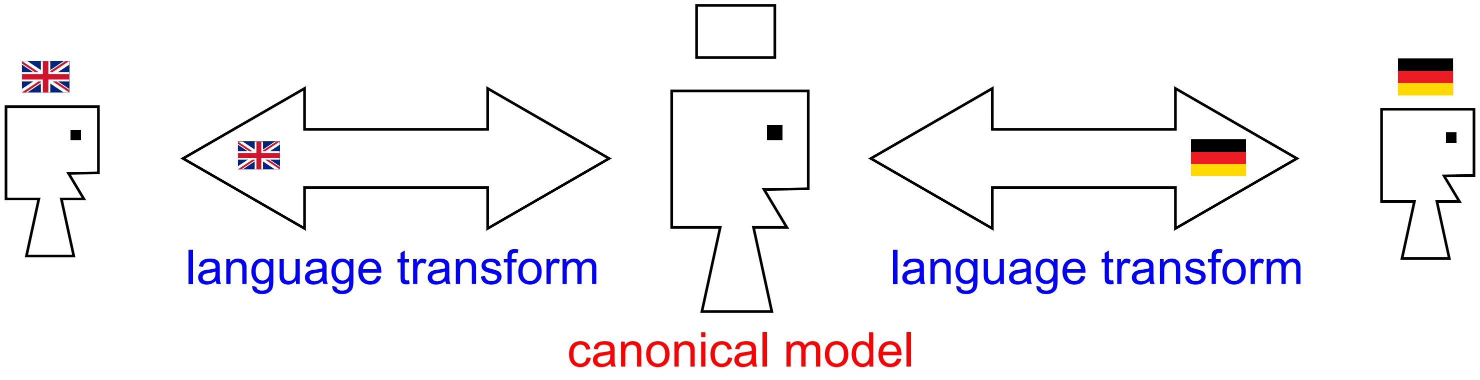
Speaker transform

- Speaker-specific characteristics

* Vocal tract length & shape, F0 height & range, voicing
* Speaking rate, speaker-specific speaking styles

⇒ Constrained MLLR [Gales;'98]

Speaker & Language Factorization (SLF)



Language transform

* Language-dependent parts of syntactic, morphological, intonational, phonetic, & phonological factors

Canonical model

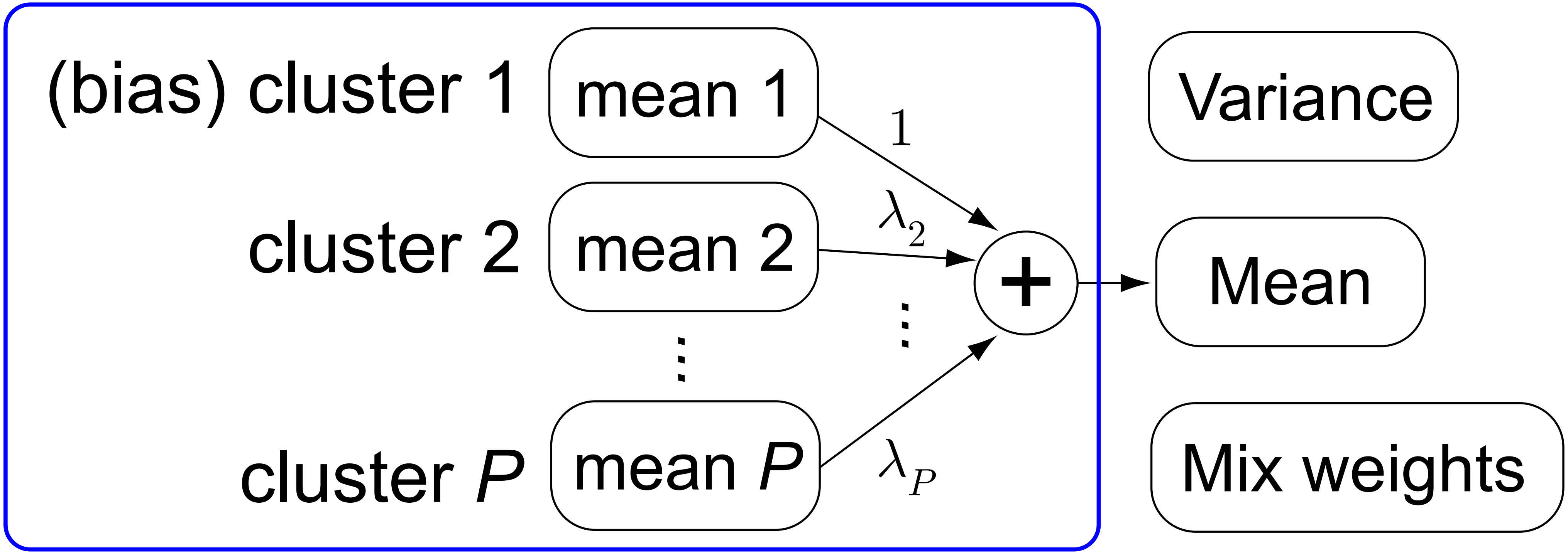
* Cross-language parts of syntactic, morphological, intonational, phonetic, & phonological factors

⇒ CAT with cluster-dependent decision trees [Zen;'09]

Cluster Adaptive Training (CAT)

Speaker adaptation by CAT [Gales;00]

- "Soft" version of speaker clustering



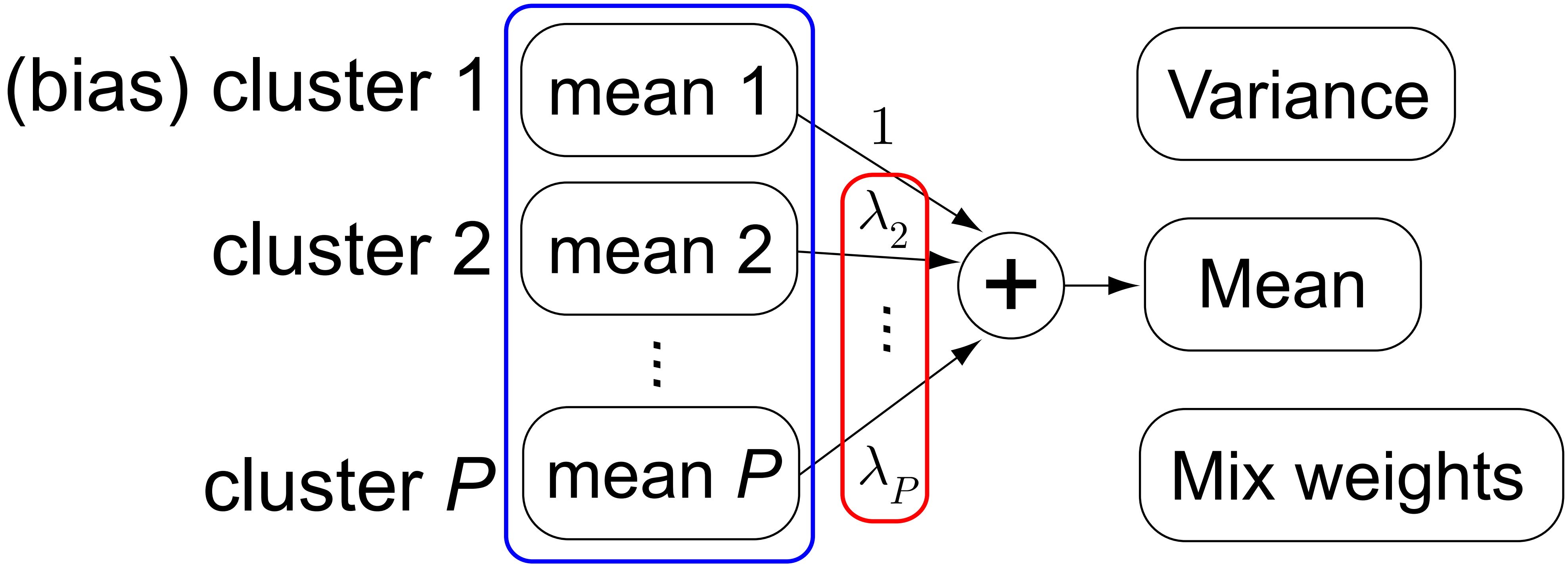
Target speaker

$\Rightarrow$ Weighted sum of underlying *prototype* speakers

Cluster Adaptive Training (CAT)

Speaker adaptation by CAT [Gales;00]

- "Soft" version of speaker clustering

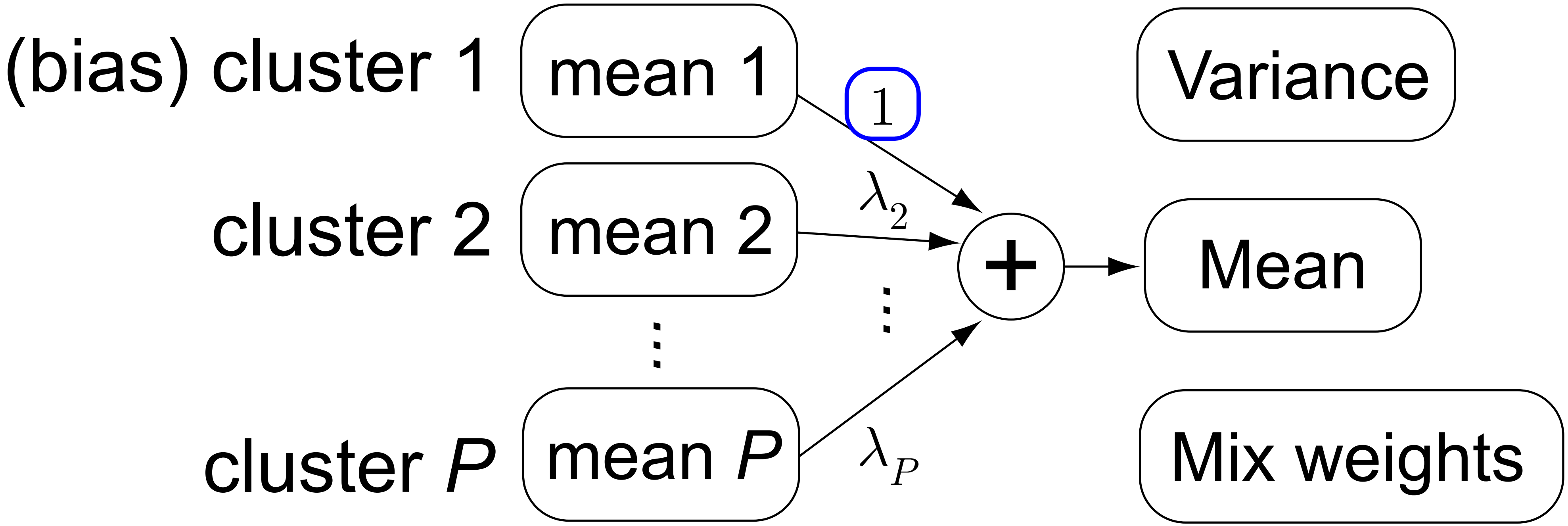


Prototype spekers are *fixed* across all speakers
Interpolation weights *change* speaker-by-speaker

Cluster Adaptive Training (CAT)

Speaker adaptation by CAT [Gales;00]

- "Soft" version of speaker clustering



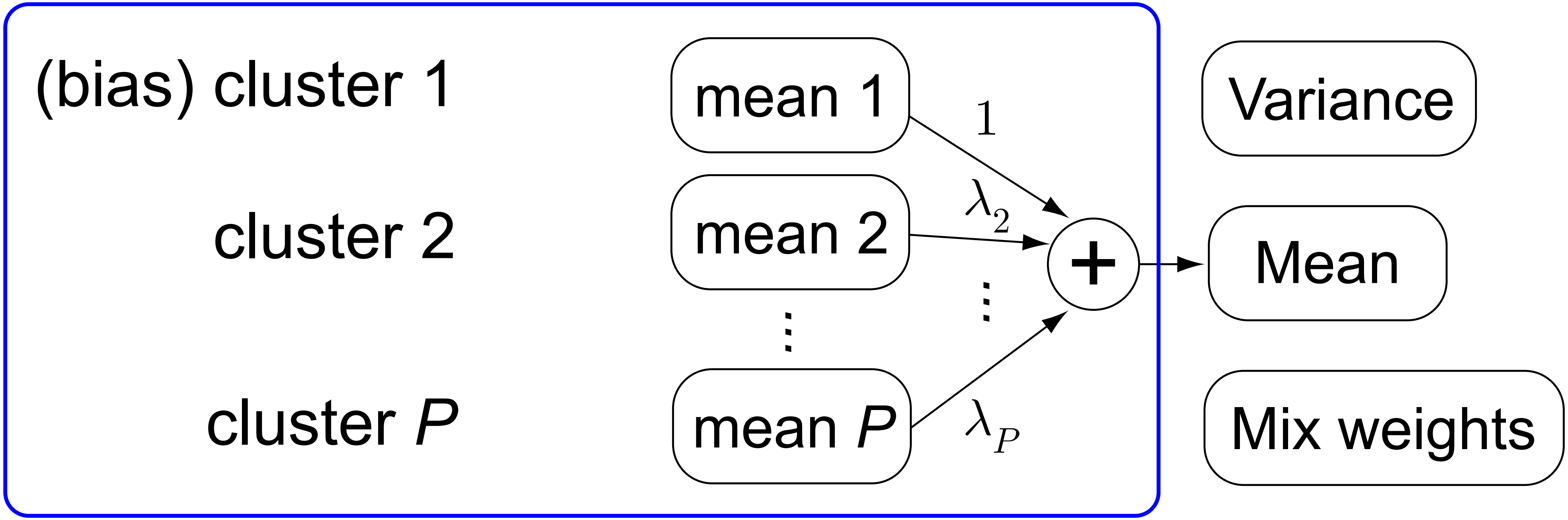
Weight for bias cluster is always equal to 1

$\Rightarrow$ Represent *common factor* across speakers

Cluster Adaptive Training (CAT)

Language adaptation by CAT

Extend CAT idea to represent languages



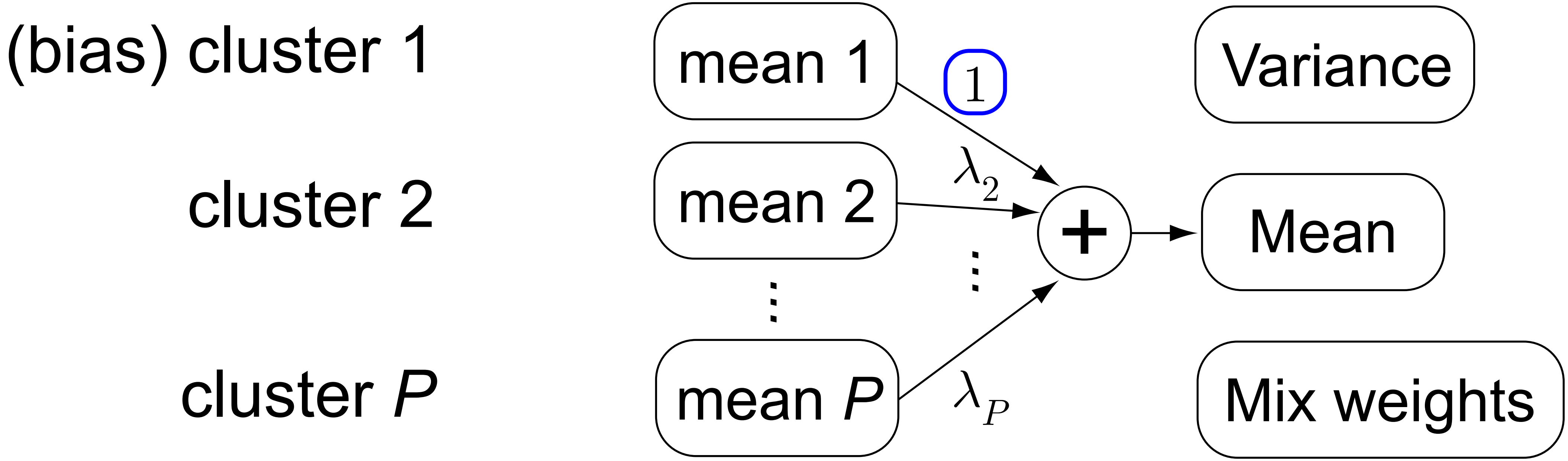
Target *language*

$\Rightarrow$ Weighted sum of underlying *prototype languages*

Cluster Adaptive Training (CAT)

Language adaptation by CAT

Extend CAT idea to represent languages



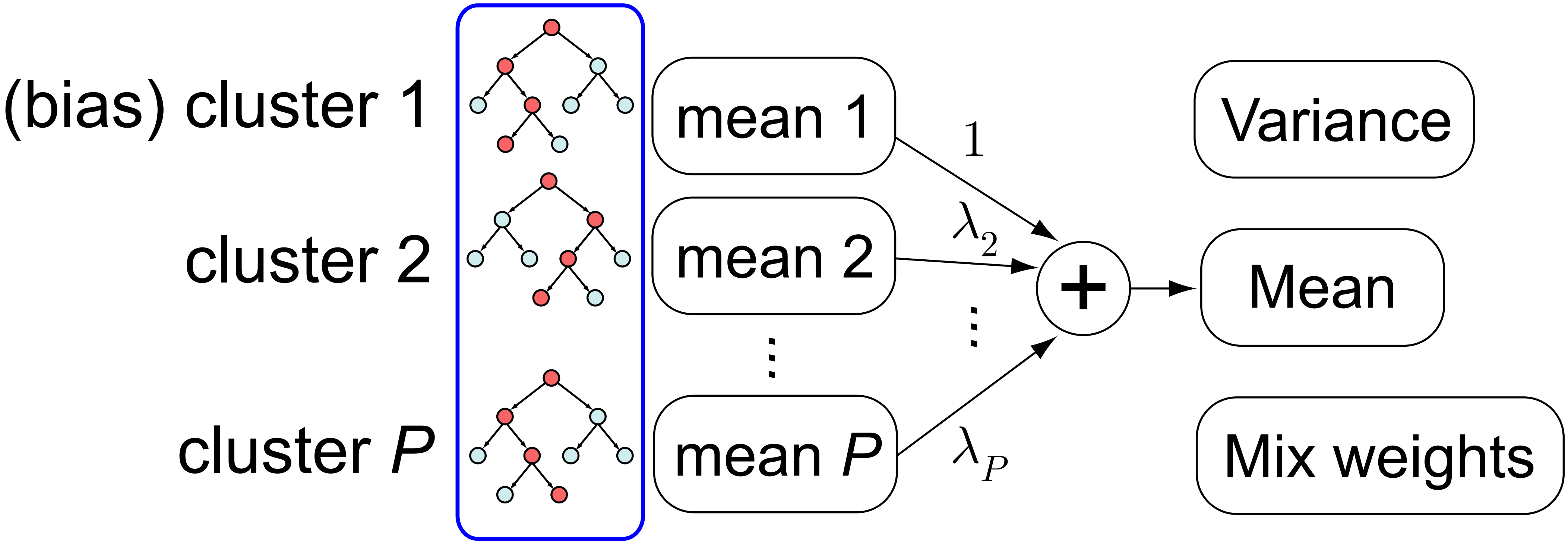
Weight for bias cluster is always equal to 1

$\Rightarrow$ Represent common factor across *languages*

Cluster Adaptive Training (CAT)

Language adaptation by CAT

Extend CAT idea to represent languages



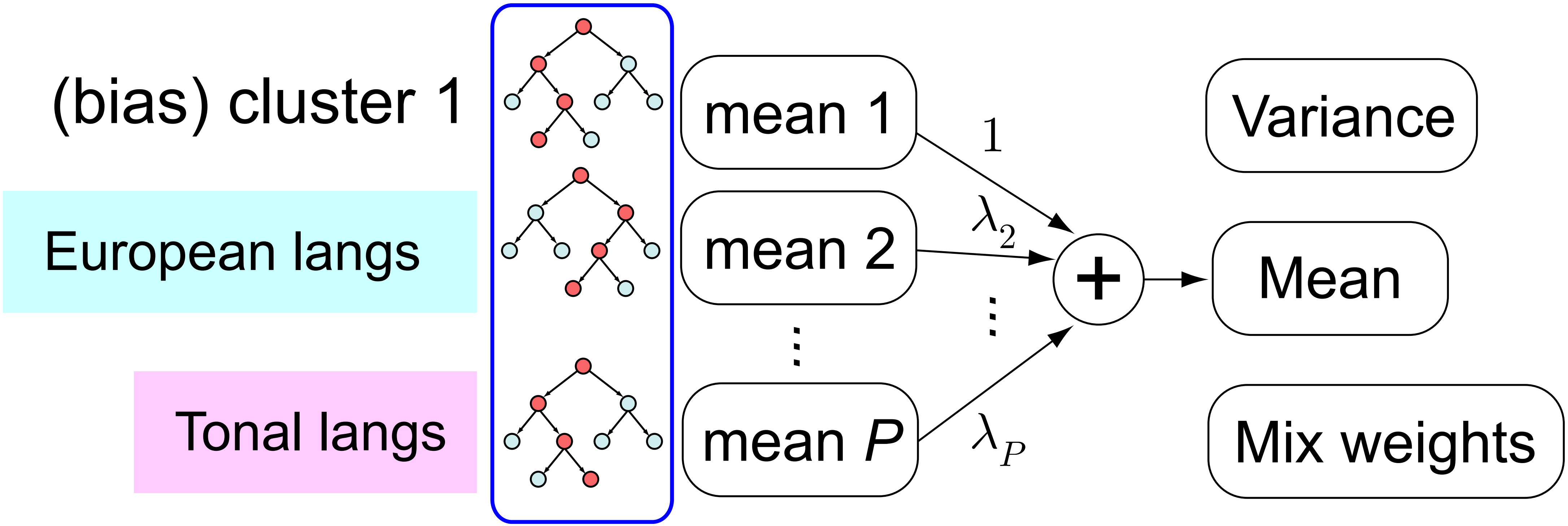
Prototype languages have their own context dependencies

$\Rightarrow$ CAT with cluster-dependent decision trees [Zen;'09]

Cluster Adaptive Training (CAT)

Language adaptation by CAT

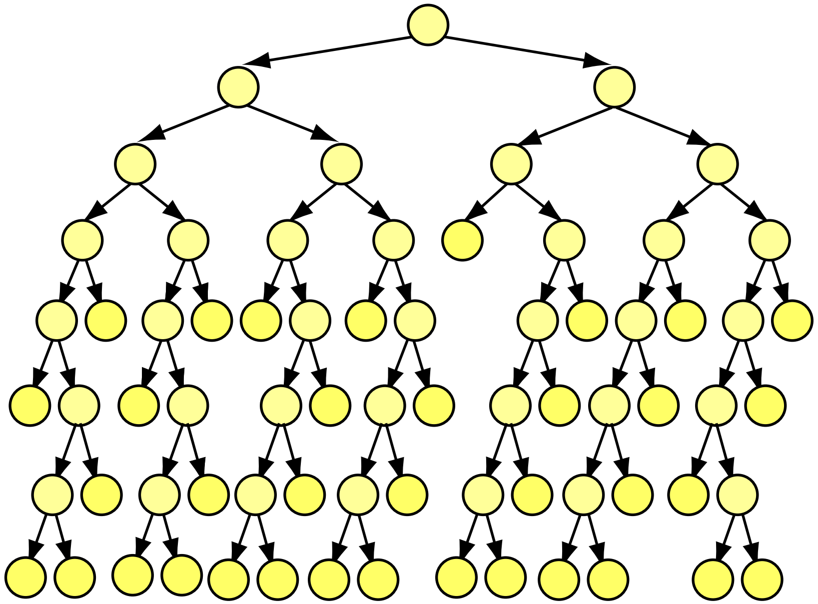
Extend CAT idea to represent languages

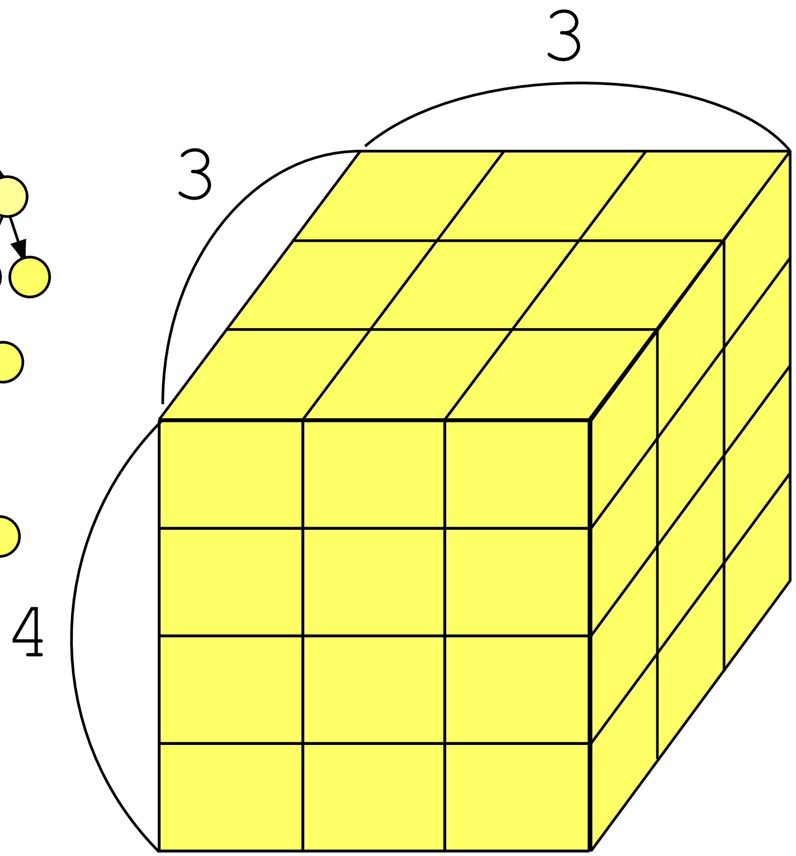


Prototype languages have their own context dependencies

$\Rightarrow$ CAT with cluster-dependent decision trees [Zen;'09]

Tree Interesection Interpretation



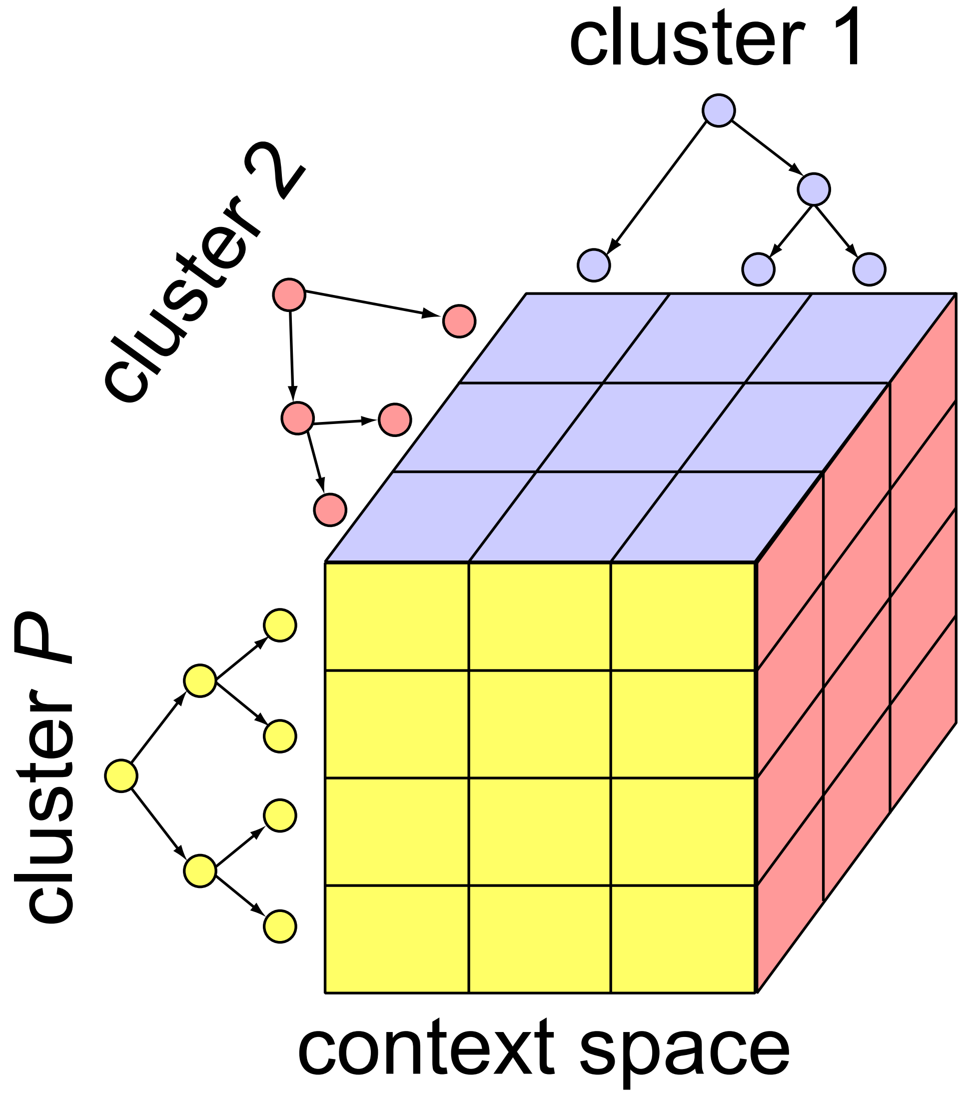


context space

3*3*4=36

#leaf nodes=36

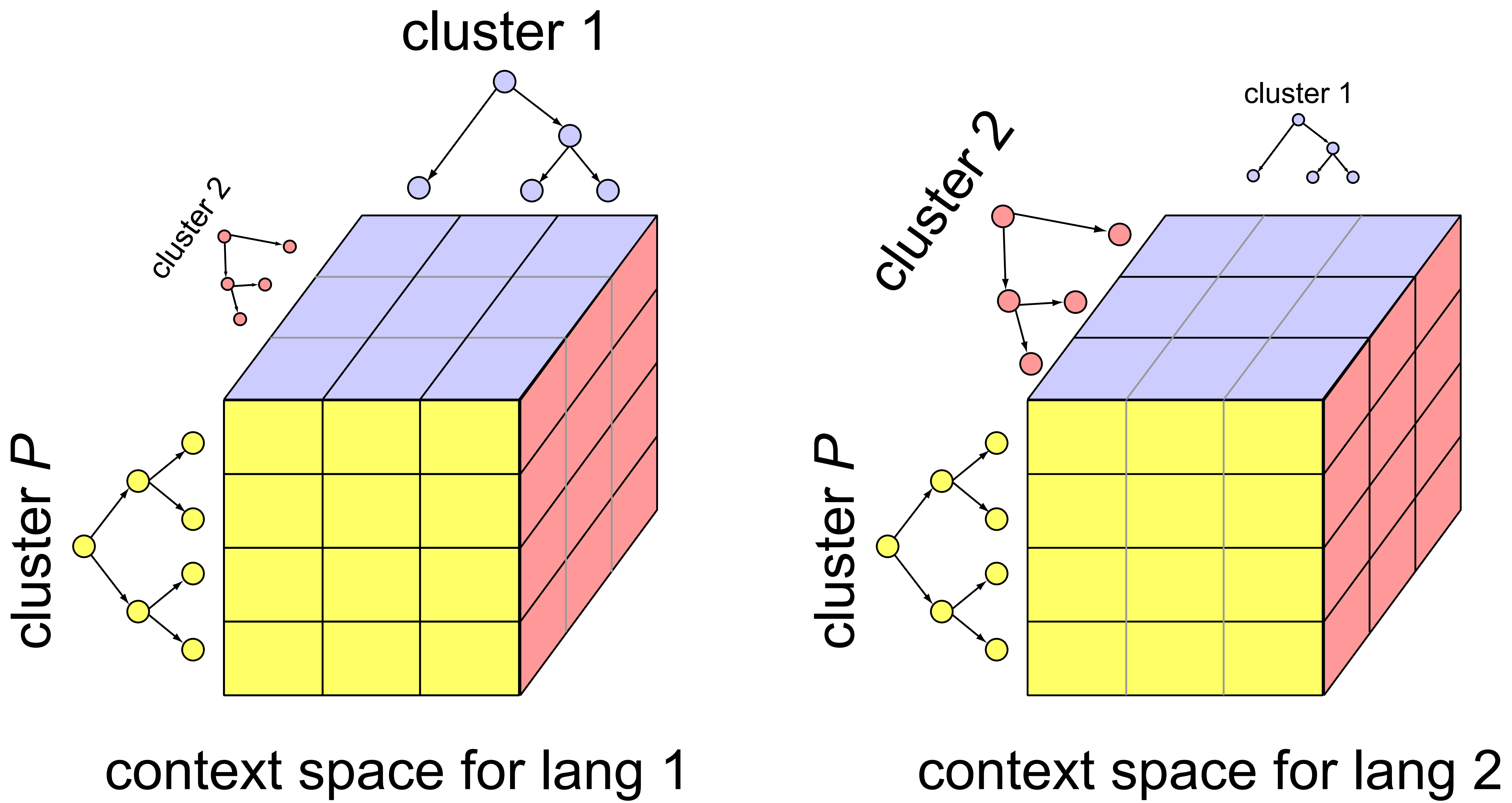
Tree Interesection Interpretation



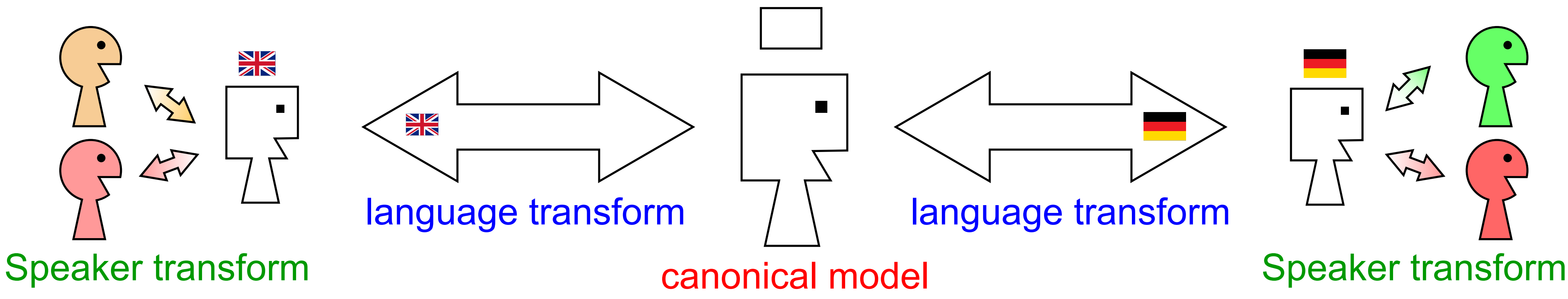
3*3*4=36

#leaf nodes=10

Tree Interesection Interpretation



Speaker & Language Factorization (SLF)



Speaker transform

⇒ CMLLR

Language transform

⇒ CAT non-bias clusters & CAT interpolation weights

Canonical model

⇒ CAT bias cluster

Trees & params can be updated iteratively by EM

Definition of State-Output Distributions

$$p(\boldsymbol{o}(t) \mid m, s, l, \mathcal{M}) = \left|\boldsymbol{A}_{r(m)}^{(s)}\right| \mathcal{N}\left(\boxed{\boldsymbol{A}_{r(m)}^{(s)} \boldsymbol{o}(t) + \boldsymbol{b}_{r(m)}^{(s)}} ; \boxed{\sum_{i=1}^{P} \lambda_{i,q(m)}^{(l)} \boldsymbol{\mu}_{c(m,i)}}, \boldsymbol{\Sigma}_{v(m)}\right)$$

CMLLR CAT

$\boldsymbol{o}(t)$: observation vector at frame t

m : mixture component index

s : speaker label associated with $\boldsymbol{o}(t)$

l : language label associated with $\boldsymbol{o}(t)$

$\boldsymbol{A}$,$\boldsymbol{b}$: CMLLR transforms

λ : CAT interpolation weights

$\boldsymbol{\mu}$: CAT cluster mean vectors

$\boldsymbol{\Sigma}$: canonical covariance matrices

$r(m)$: CMLLR regression class

$q(m)$: CAT regression class

$c(m,i)$: mean vector index

$v(m)$: covariance matrix index

Training Process

ML estimation by EM algorithm

- **Iteratively re-estimate trees, CAT & CMLLR params**
- **Training process**

  1) Initialize trees, CAT & CMLLR params

  2) Re-construct trees

  3) Re-estimate CAT params while fixing CMLLR params

  4) Re-estimate CMLLR params while fixing CAT params

  5) Go to 2) until converge

Estimation

Update formulae

- **CMLLR transform**
  * Same as normal CMLLR estimation [Gales;'98]
- **CAT weights**
  * Same as normal CAT estimation [Gales;'00]
- **Canonical covariance matrices & mixture weights**
  * Straightforward
- **Canonical cluster mean vectors**
  * All cluster mean vectors depend on each other due to trees
  * Trees are iteratively reconstructed

Update Formulae of SLF Cluster Mean Vectors

Auxiliary function

$$\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}) = -\frac{1}{2}\sum_{m,i}\Bigg(\boldsymbol{\mu}_{c(m,i)}^{\top}\boldsymbol{G}_{ii}^{(m)}\boldsymbol{\mu}_{c(m,i)} + 2\sum_{j\neq i}\boldsymbol{\mu}_{c(m,i)}^{\top}\boldsymbol{G}_{ij}^{(m)}\boldsymbol{\mu}_{c(m,j)} - 2\boldsymbol{\mu}_{c(m,i)}^{\top}\boldsymbol{k}_{i}^{(m)}\Bigg)$$

$$\boldsymbol{G}_{ij}^{(m)} = \sum_{t,l}\gamma_m(t)\lambda_{i,q(m)}^{(l)}\boldsymbol{\Sigma}_{v(m)}^{-1}\lambda_{j,q(m)}^{(l)}$$

$$\boldsymbol{k}_{i}^{(m)} = \sum_{t,s,l}\gamma_m(t)\lambda_{i,q(m)}^{(l)}\boldsymbol{\Sigma}_{v(m)}^{-1}\underline{\hat{\boldsymbol{o}}_{r(m)}^{(s)}(t)}$$

CMLLR-transformed
observation vector

Update Formulae of SLF Cluster Mean Vectors

Derivative of auxiliary function

$$\boldsymbol{G}_{n\nu} = \sum_{\substack{m,i,j \\ c(m,i)=n \\ c(m,j)=\nu}} \boldsymbol{G}_{ij}^{(m)} \qquad \boldsymbol{k}_n = \sum_{\substack{m,i \\ c(m,i)=n}} \boldsymbol{k}_i^{(m)}$$

$$\frac{\partial \mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}})}{\partial \boldsymbol{\mu}_n} = \boldsymbol{k}_n - \boldsymbol{G}_{nn}\boldsymbol{\mu}_n - \sum_{\nu \neq n} \boldsymbol{G}_{n\nu}\boldsymbol{\mu}_\nu \quad \Rightarrow \mathbf{0}$$

$$\hat{\boldsymbol{\mu}}_n = \boldsymbol{G}_{nn}^{-1} \left(\boldsymbol{k}_n - \sum_{\nu \neq n} \boldsymbol{G}_{n\nu} \underline{\boldsymbol{\mu}_\nu} \right)$$

ML estimate of a CAT mean vector

$\Rightarrow$ **depends on all the other CAT mean vectors**

Update Formulae of SLF Cluster Mean Vectors

Joint update of all cluster mean vectors

$$\begin{bmatrix} \boldsymbol{G}_{11} & \cdots & \boldsymbol{G}_{1N} \\ \vdots & \ddots & \vdots \\ \boldsymbol{G}_{N1} & \cdots & \boldsymbol{G}_{NN} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}}_1 \\ \vdots \\ \hat{\boldsymbol{\mu}}_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{k}_1 \\ \vdots \\ \boldsymbol{k}_N \end{bmatrix}$$

$$\boldsymbol{G}_{n\nu} = \sum_{\substack{m,i,j,t,l \\ c(m,i)=n \\ c(m,j)=\nu}} \gamma_m(t) \lambda_{i,q(m)}^{(l)} \boldsymbol{\Sigma}_{v(m)}^{-1} \lambda_{j,q(m)}^{(l)} \qquad \boldsymbol{k}_n = \sum_{\substack{m,i,t,s,l \\ c(m,i)=n}} \gamma_m(t) \lambda_{i,q(m)}^{(l)} \boldsymbol{\Sigma}_{v(m)}^{-1} \underline{\hat{\boldsymbol{o}}_{r(m)}^{(s)}(t)}$$

transformed
observation

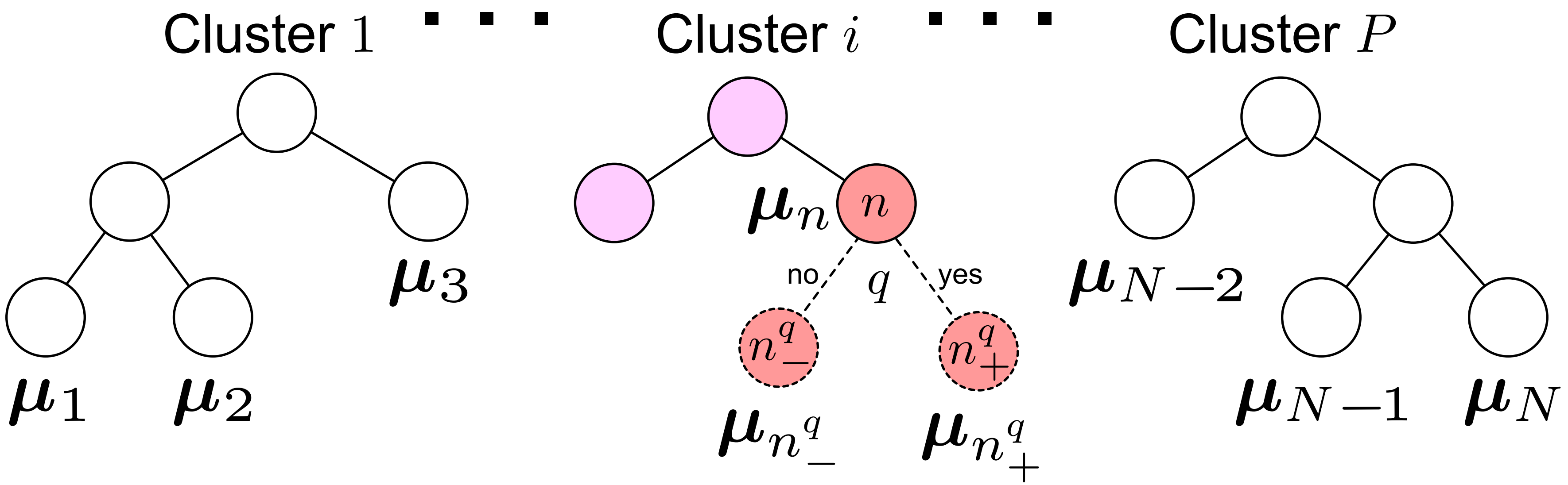
Size of linear equations > 10,000, but sparse

$\Rightarrow$ Sparse storage (CSR) & solver (CG or PARDISO)

All CAT mean vectors can be determined jointly

Update Procedure of Decision Trees

Rebuild tree while fixing other trees & params

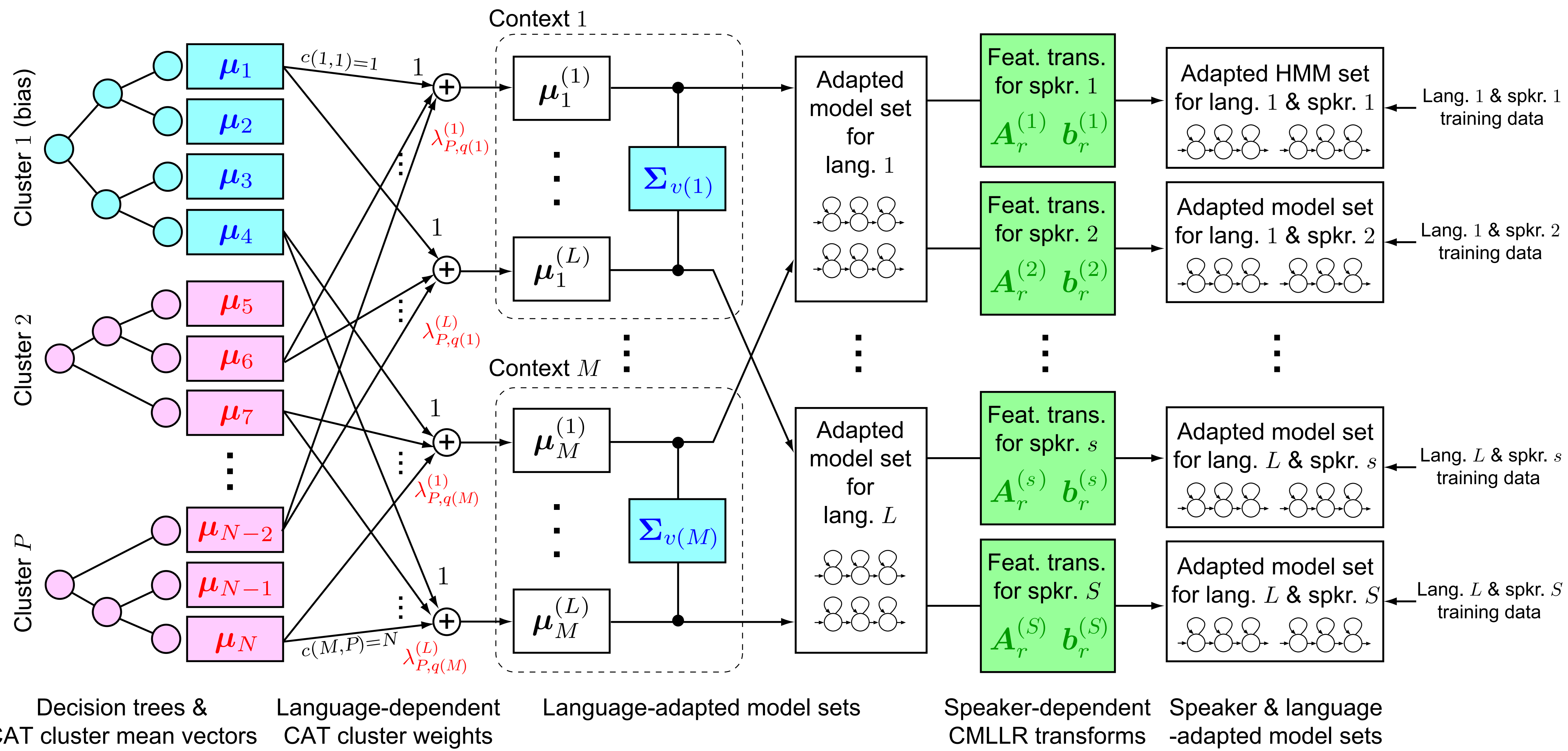


Log likelihood

$$\mathcal{L}(n) = \frac{1}{2} \sum_{m \in S(n)} \left(\boldsymbol{k}_i^{(m)} - \sum_{j \neq i} \boldsymbol{G}_{ij}^{(m)} \boldsymbol{\mu}_{c(m,j)} \right)^\top \left(\sum_{m \in S(n)} \boldsymbol{G}_{ii}^{(m)} \right)^{-1} \sum_{m \in S(n)} \left(\boldsymbol{k}_i^{(m)} - \sum_{j \neq i} \boldsymbol{G}_{ij}^{(m)} \boldsymbol{\mu}_{c(m,j)} \right)$$

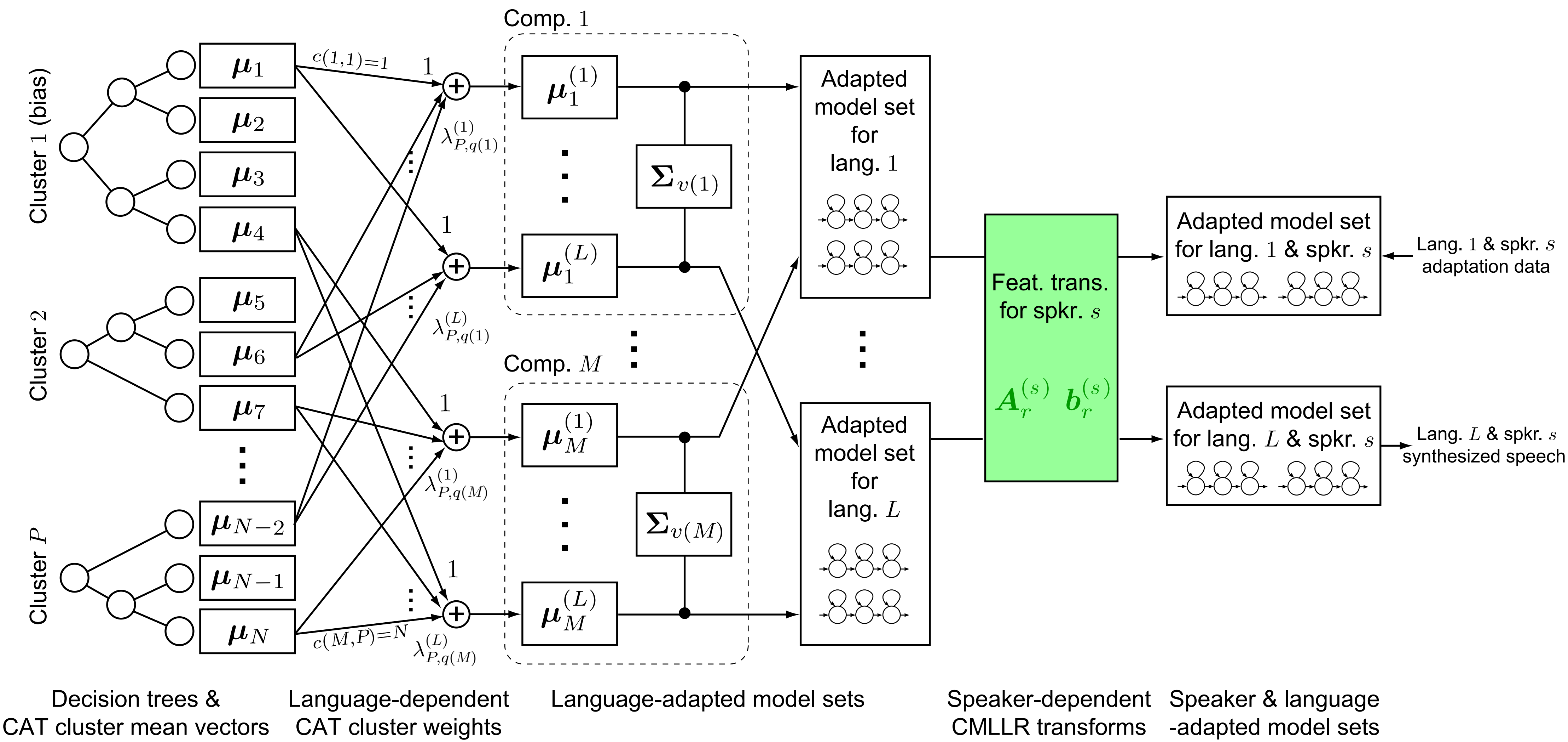
$\longrightarrow$ Trees can be updated one-by-one

Block Diagram of SLF Training

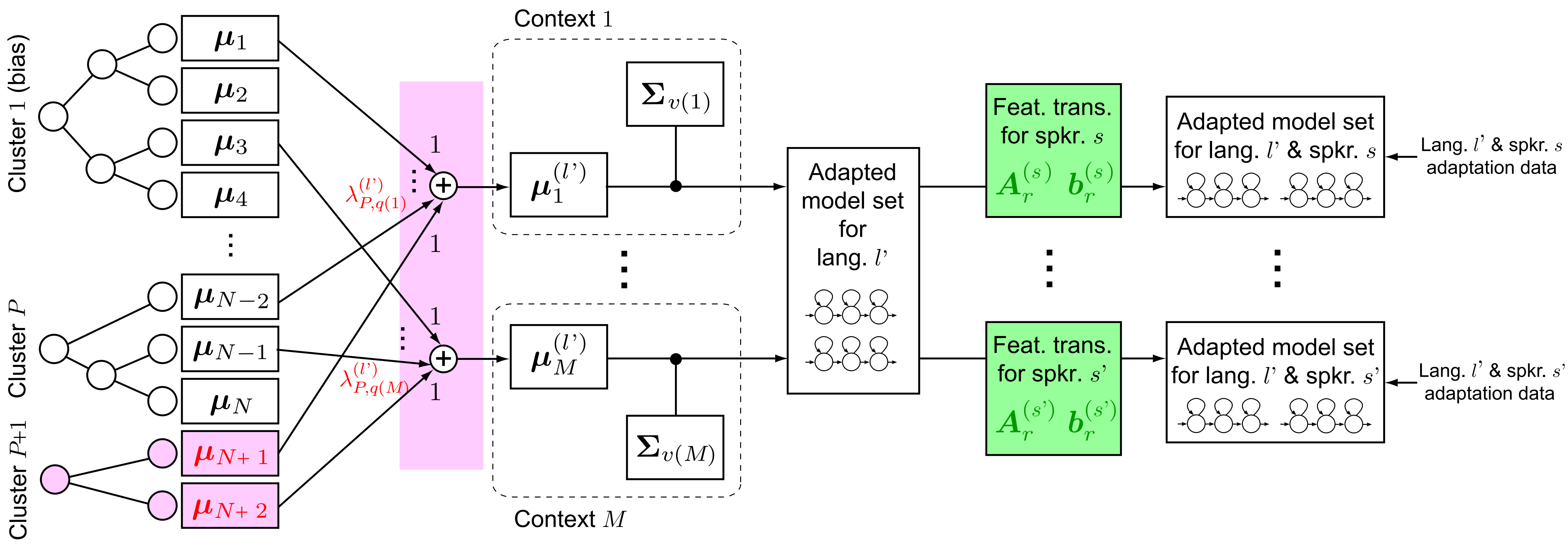


canonical model | language transform | speaker transform

Block Diagram of SLF Cross-Lingual Adaptation



Block Diagram of SLF Language Adaptation



Decision trees & CAT cluster mean vectors

Language-dependent CAT cluster weights

Language-adapted model sets

Speaker-dependent CMLLR transforms

Speaker & language -adapted model sets

Outline

- Background
- Conventional approaches
    * Polyglot speaker
    * Mixing mono-lingual corpora
    * Cross-lingual speaker adaptation
- Speaker & language factorization (SLF)
    * Concept
    * Details
- Experiments
- Conclusions

Experimental Conditions

Data

- German, French, Spanish, UK & US English
- 10 speakers per language (5 female & 5 male)
  - 8 speakers for training, 2 speakers for adaptation & test
- 100~150 utterances per speaker
- Consistent microphone & recording condition

Data preparation

- IPA-like universal phone set
- Universal context-dependent label format
  * phone, syllable, word, phrase, & utterance-level contexts

Experimental Conditions

Speech analysis / training / synthesis setup

- Similar to HTS-2008 (SAT system for BC08) [Yamagishi;'08]
  * 39 mel-cepstrum, log F0, 23 Bark critical band aperiodicity
  * Delta & Delta-Delta
- LI-SAT (language-independent) was trained
- Initialize SLF model by LI-SAT model then reestimate
- LD-SAT (language-dependent) models were also trained
- Cov mats & mix weights had the same tree as bias cluster
- 3 regression classes for CAT & CMLLR
  * silence, short pause, & speech
- Speech parameter generation algorithm with GV [Toda;'07]

Number of Leaf Nodes

Cluster	mel-cep	log F0	band ap	dur
1 (bias)	2,071	4,059	5,940	1,168
2	102	3,304	20	46
3	164	3,744	17	38
4	88	3,582	18	27
5	129	3,259	25	21
6	125	2,956	28	41
Total	2,679	20,904	6,048	1,341
LI-SAT	2,235	7,557	6,014	1,371
LD-SAT	2,957	9,129	6,551	1,739

Total sizes of trees were comparable

Number of Leaf Nodes

Cluster	mel-cep	log F0	band ap	dur
1 (bias)	2,071	4,059	5,940	1,168
2	102	3,304	20	46
3	164	3,744	17	38
4	88	3,582	18	27
5	129	3,259	25	21
6	125	2,956	28	41
Total	2,679	20,904	6,048	1,341

Bias cluster was largest in all speech params

⇒ Common factor across languages was dominant

Number of Leaf Nodes

Cluster	mel-cep	log F0	band ap	dur
1 (bias)	2,071	4,059	5,940	1,168
2	102	3,304	20	46
3	164	3,744	17	38
4	88	3,582	18	27
5	129	3,259	25	21
6	125	2,956	28	41
Total	2,679	20,904	6,048	1,341

Non-bias clusters had large number of leaf nodes

$\Rightarrow$ Language-dependent factors had large contribution

Examples of CAT Interpolation Weights

mel-cep

	1	2	3	4	5	6
German	[1	0.62	.40	-0.02	.34	.33]
UK English	[1	.29	.58	.42	.25	.23]
US English	[1	.34	.46	.85	.26	.24]
Spanish	[1	.49	.38	.05	.63	.40]
French	[1	.43	.31	-0.07	.38	.68]

log F0

	1	2	3	4	5	6
German	[1	0.90	.05	.14	.10	.10]
UK English	[1	.04	.88	.18	.06	.08]
US English	[1	.11	.20	.82	.04	.09]
Spanish	[1	.06	.12	.12	.91	.08]
French	[1	.06	.05	.17	.09	.91]

Paired Comparison Test

Preference test among LD-SAT, LI-SAT, & SLF

- 50 test sentences excluded from training data / language
- Carried out on Amazon Mechanical Turk

Results

Language	LD-SAT	LI-SAT	SLF	No pref.
German	39.7	36.2	–	24.1
	35.2	–	46.8	18.0
	–	33.8	43.2	23.0
US English	29.1	55.3	–	15.6
	26.2	–	60.6	13.1
	–	36.7	47.6	15.6

Evaluation of Cross-Lingual Adaptation

DMOS & MOS test setup

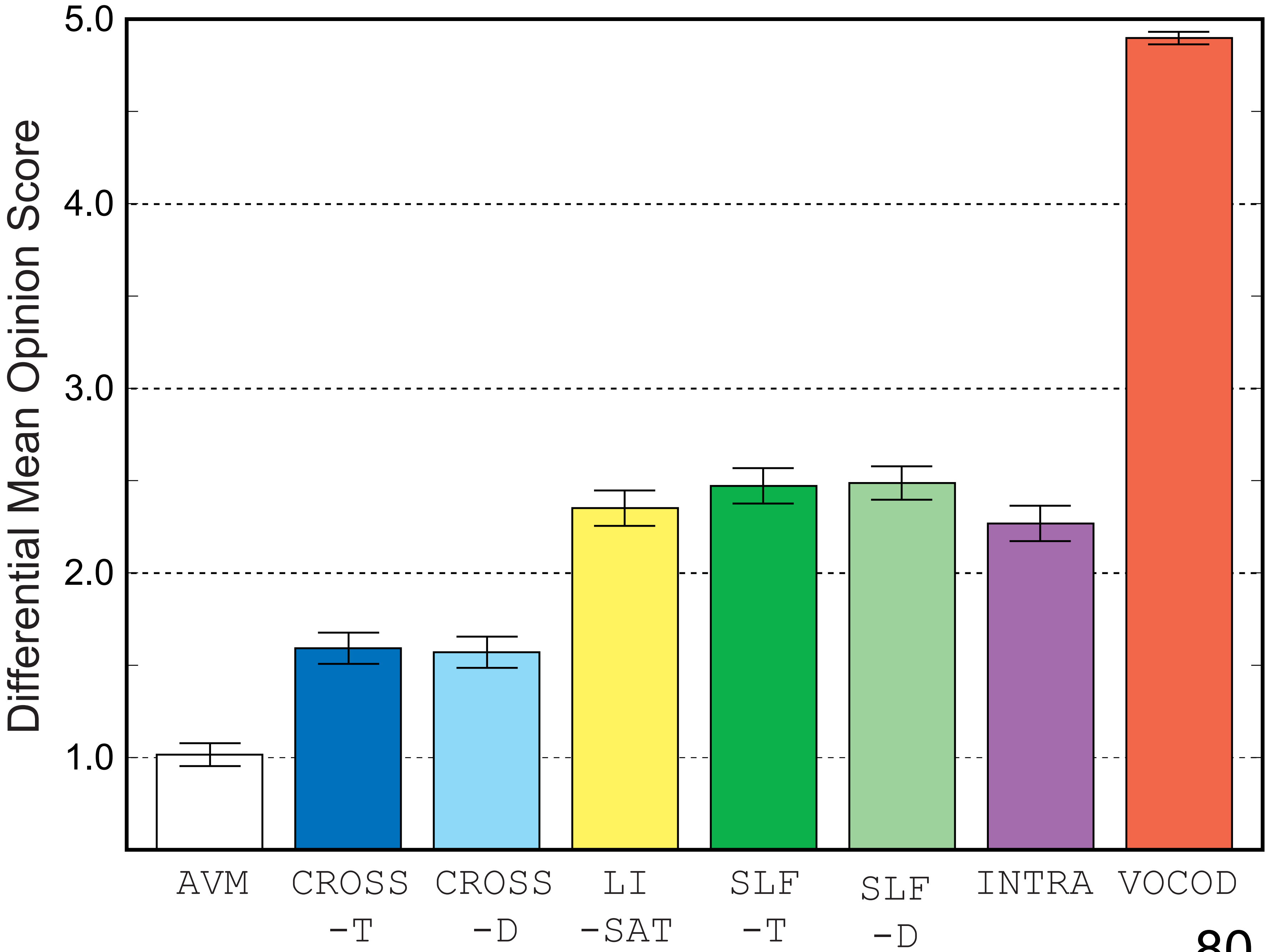
- Target speakers: 6 German speakers from EMIME
  German/English bilingual corpus
- Target language was English
- Amazon Mechanical Turk
- 5-scale similarity/naturalness score
  * DMOS 1: very dissimilar - 5: very similar
  * MOS 1: very natural - 5: very unnatural

Evaluation of Cross-Lingual Adaptation

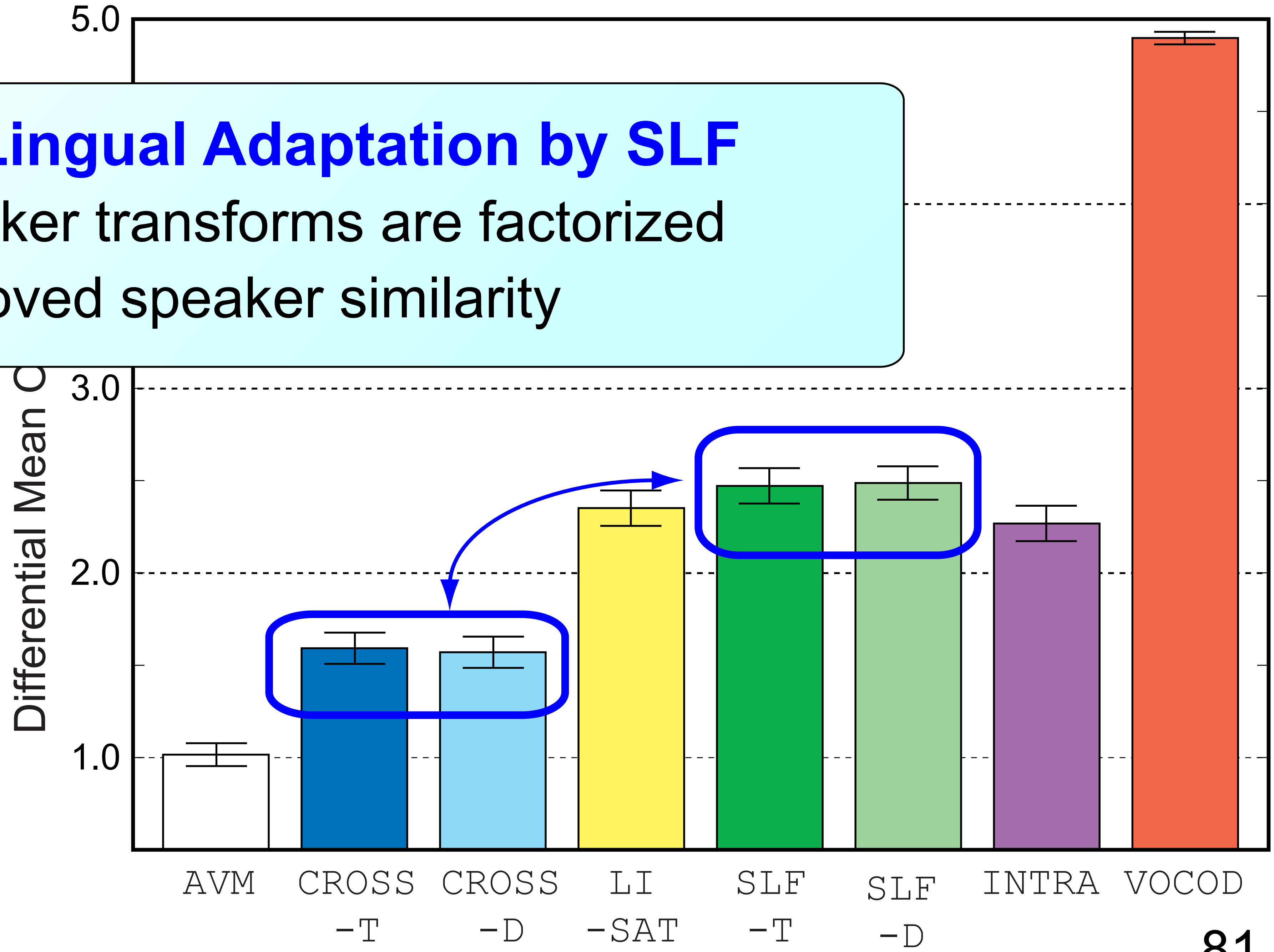
Systems to be compared

1) US English LD-SAT w/o adaptation (AVM)
2) US English LD-SAT adapted by state-mapping cross-lingual speaker adaptation based on transform mapping (CROSS-T)
3) US English LD-SAT adapted by state-mapping cross-lingual speaker adaptation based on data mapping (CROSS-D)
4) LI-SAT w/ adaptation (LI-SAT)
5) SLF adapted by transform mapping (SLF-T)
6) SLF adapted by data mapping (SLF-D)
7) US English LD-SAT adapted by targets' English data (INTRA)
8) Vocoded natural speech (VOCOD)

Speaker Similarity by Cross-Lingual Adaptation



Speaker Similarity by Cross-Lingual Adaptation



Cross-Lingual Adaptation by SLF

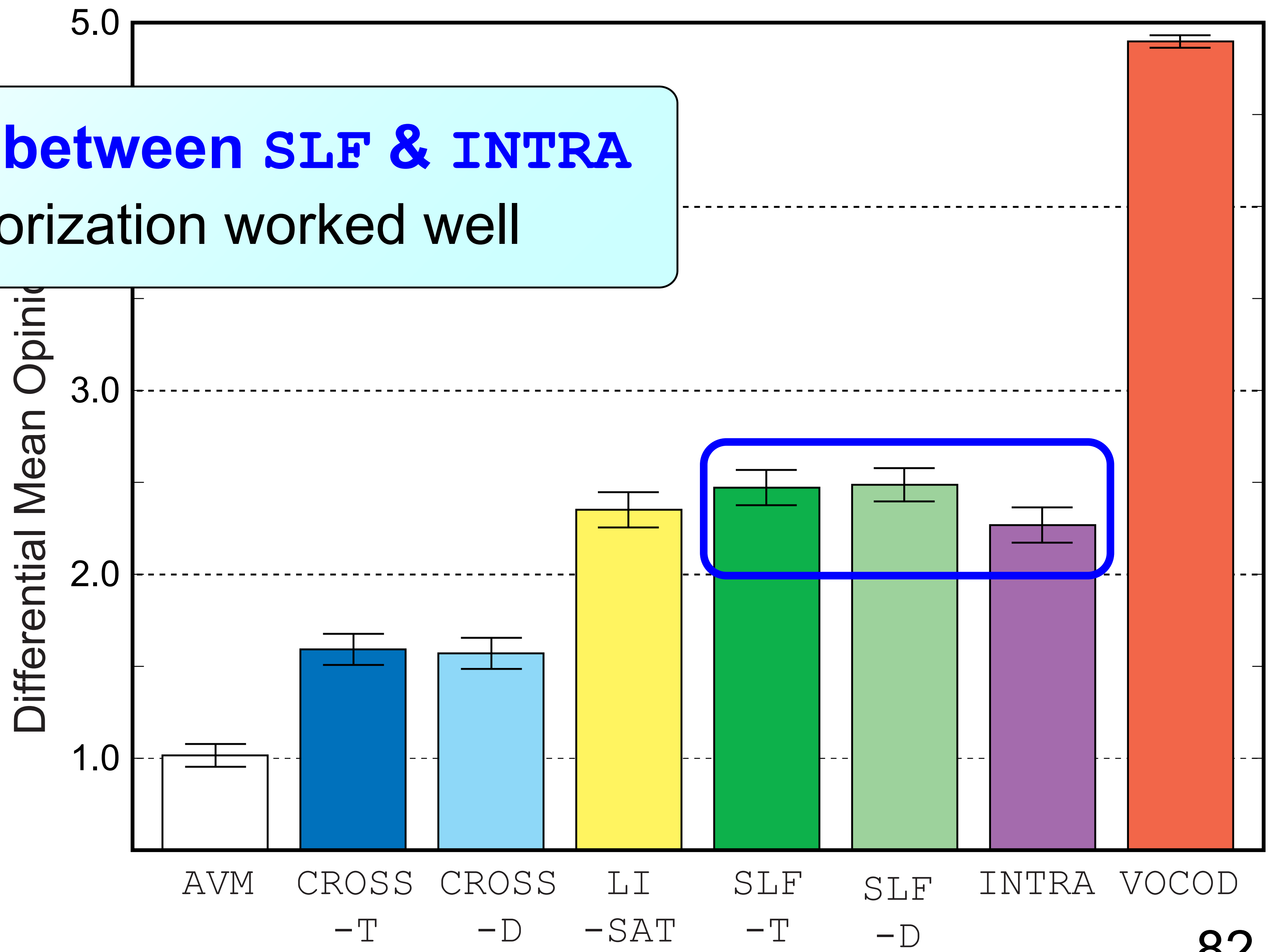
$\Rightarrow$ Speaker transforms are factorized

$\Rightarrow$ Improved speaker similarity

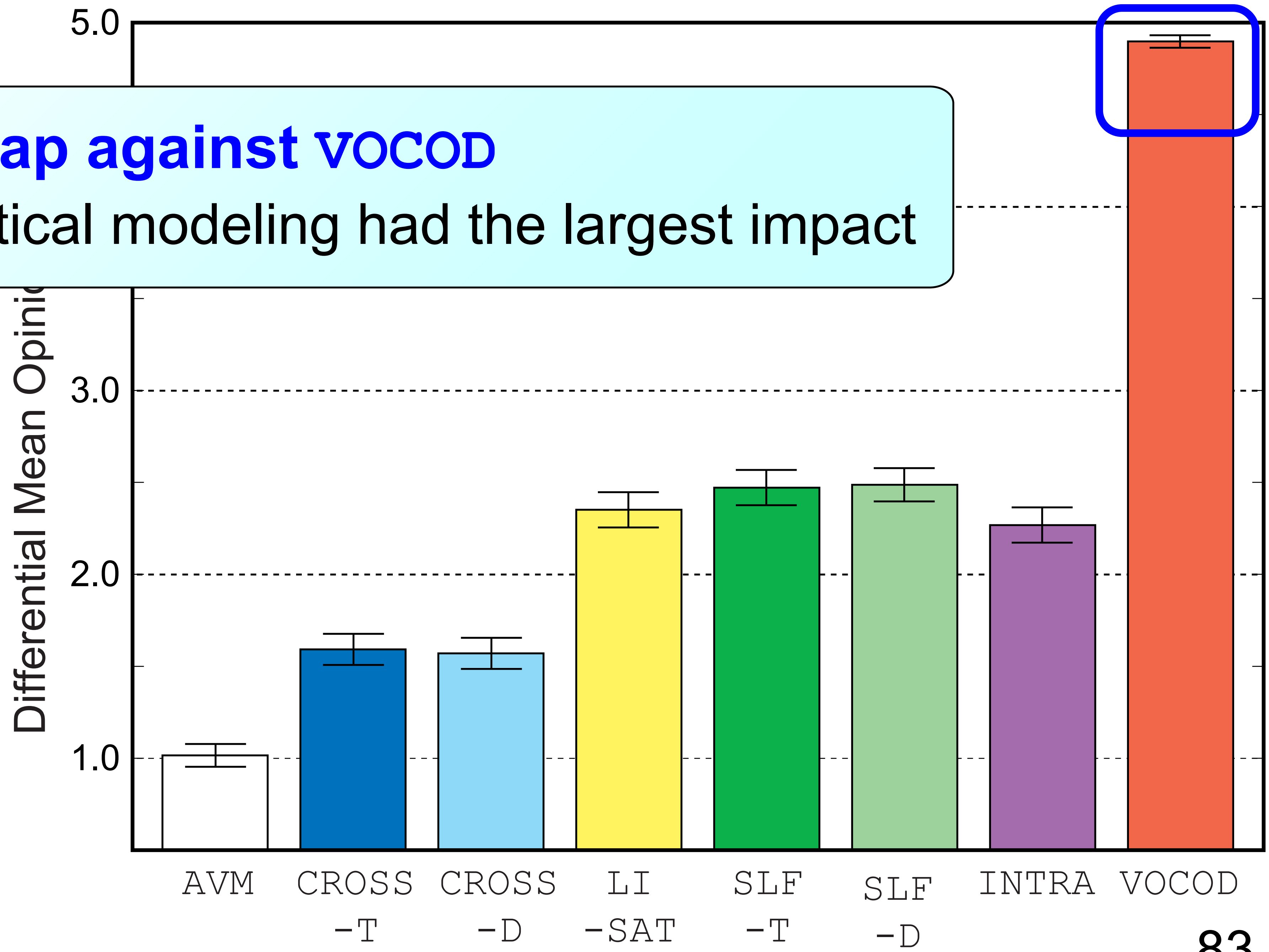
Speaker Similarity by Cross-Lingual Adaptation

No gap between `SLF` & `INTRA`

⇒ Factorization worked well



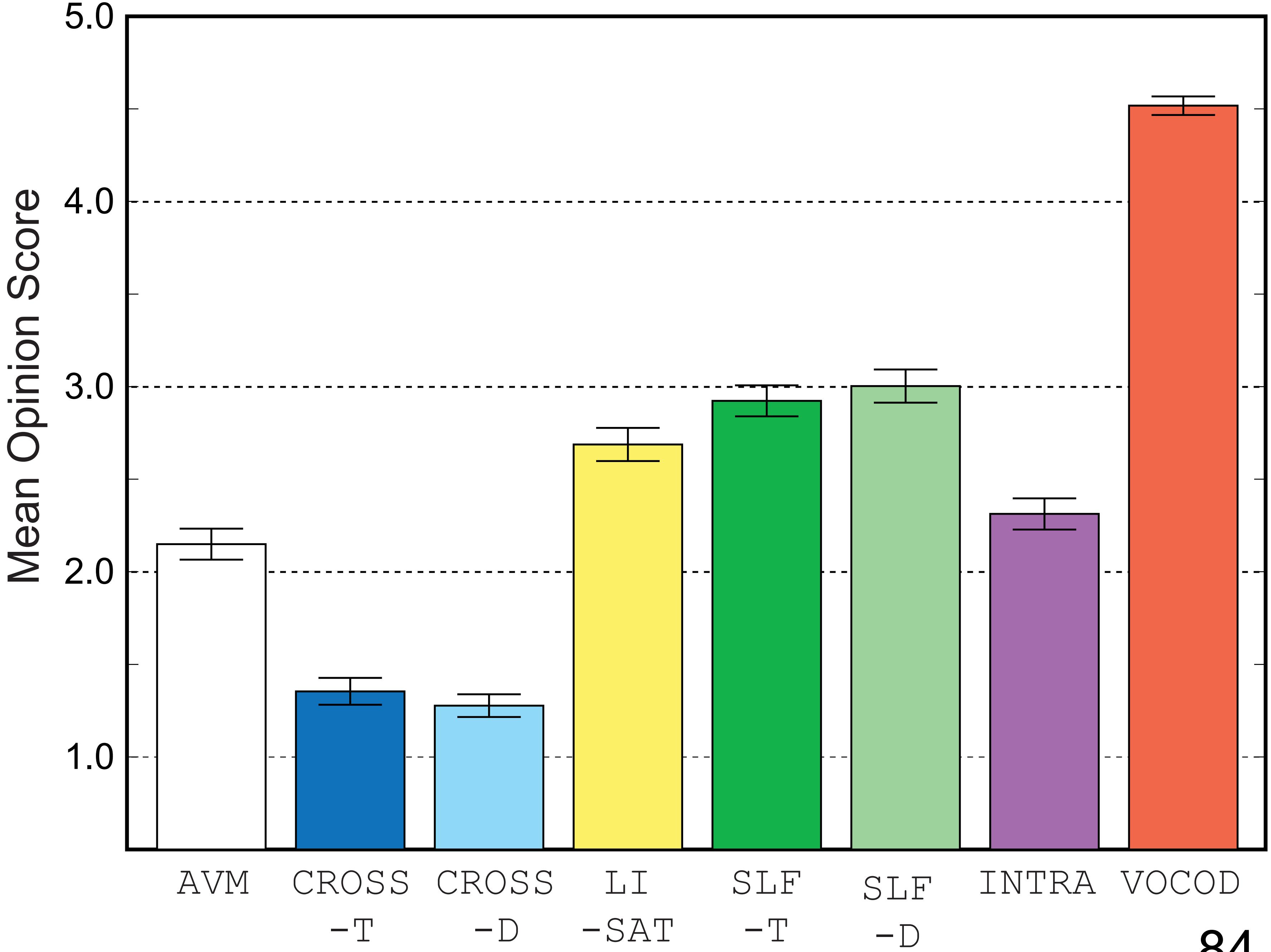
Speaker Similarity by Cross-Lingual Adaptation



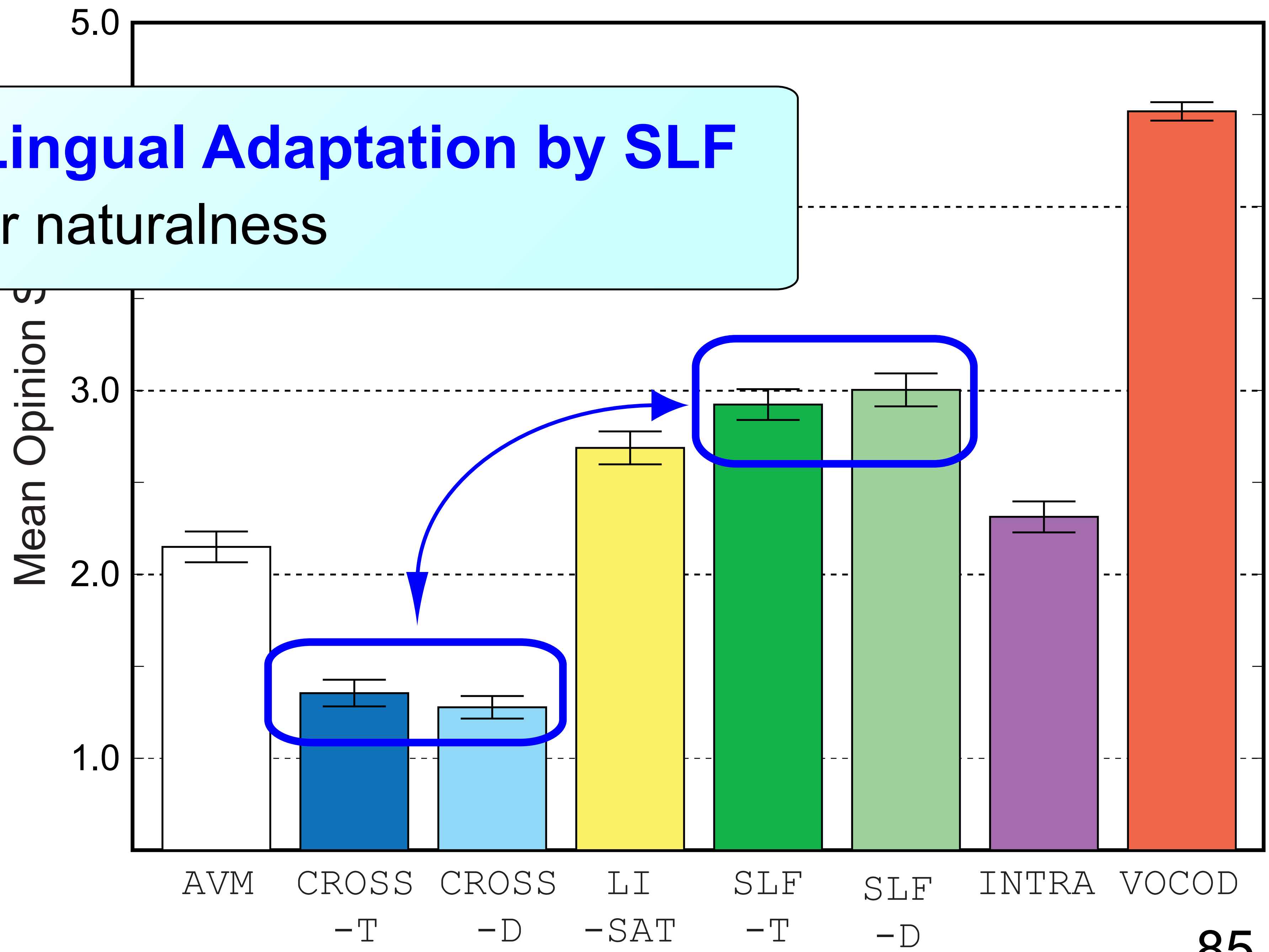
Large gap against VOCOD

⇒ Statistical modeling had the largest impact

Naturalness by Cross-Lingual Adaptation



Naturalness by Cross-Lingual Adaptation



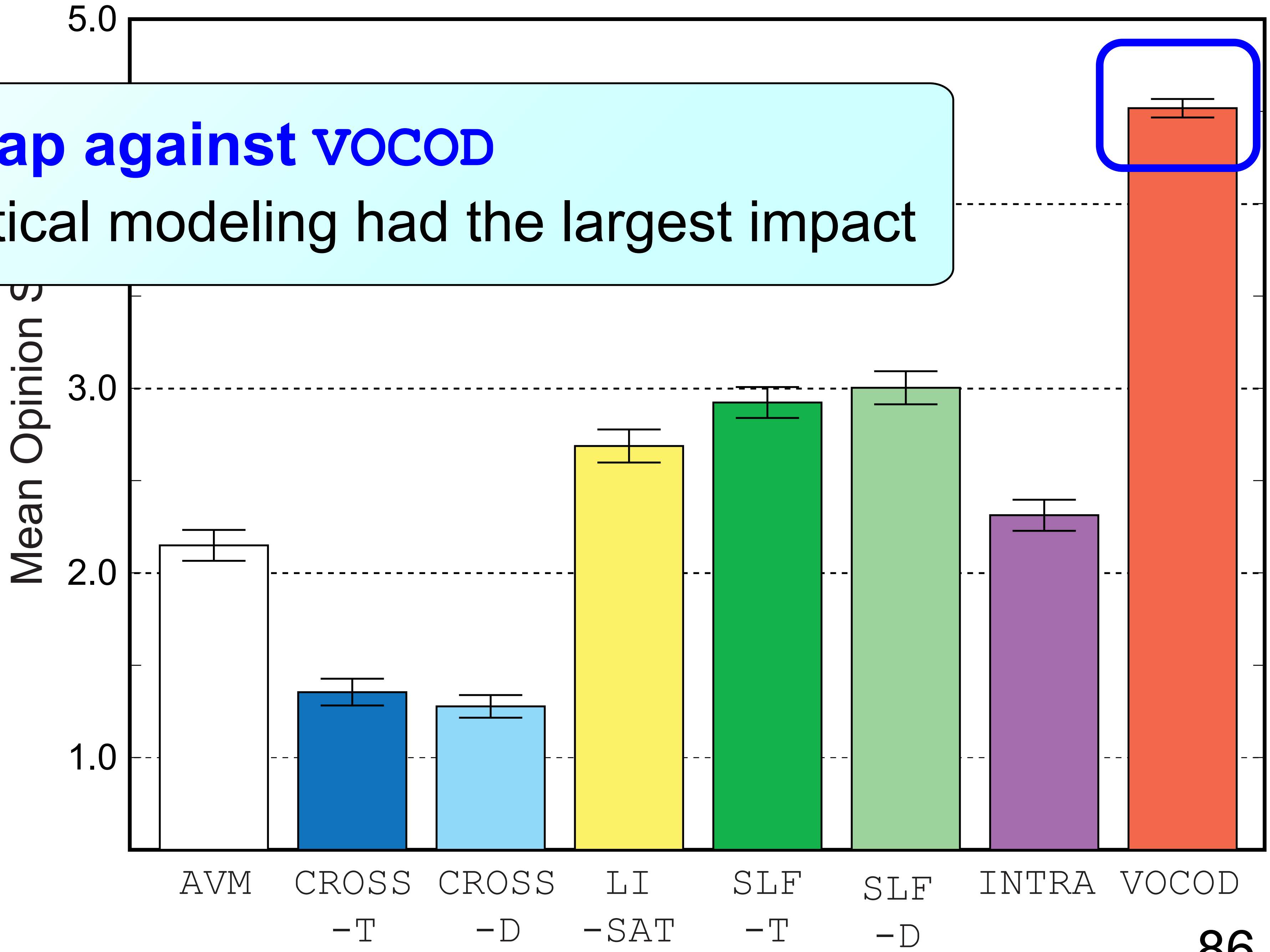
Cross-Lingual Adaptation by SLF

$\Rightarrow$ Better naturalness

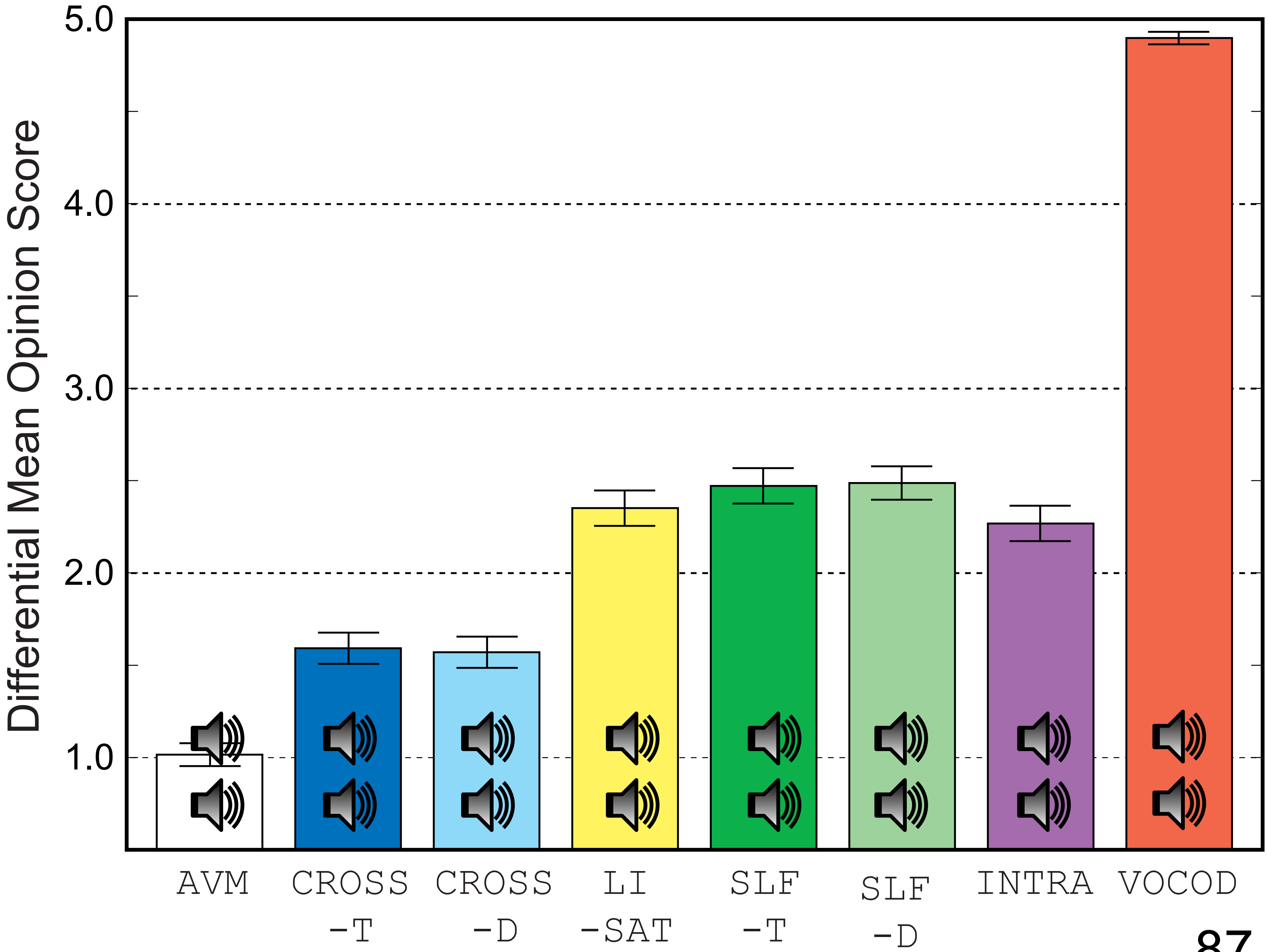
Naturalness by Cross-Lingual Adaptation

Large gap against VOCOD

⇒ Statistical modeling had the largest impact



Speaker Similarity by Cross-Lingual Adaptation



Evaluation of Language Adaptation

Experimental setup

- 1 of 5 languages was excluded from training data
- Estimate language transform
  * 8 speakers in target language
- Adapt to 2 target speakers in target language
- Amazon Mechanical Turk
- Preference test about naturalness
- 5-scale naturalness score
  (1: very natural - 5: very unnatural)

Examples of CAT Interpolation Weights

Language	Parameter	clusters 2	3	4	5
German (training)	mel-cep.	**.617**	.414	.361	.318
	log F0	**.929**	.087	.119	.084
UK English (training)	mel-cep.	.366	**.695**	.280	.274
	log F0	.040	**.914**	.060	.077
Spanish (training)	mel-cep.	.481	.374	**.645**	.414
	log F0	.061	.146	**.927**	.102
French (training)	mel-cep.	.477	.258	.411	**.712**
	log F0	.029	.119	.080	**.937**
US English (target)	mel-cep.	.362	**.535**	.273	.277
	log F0	.014	**.284**	.029	.035

Examples of CAT Interpolation Weights

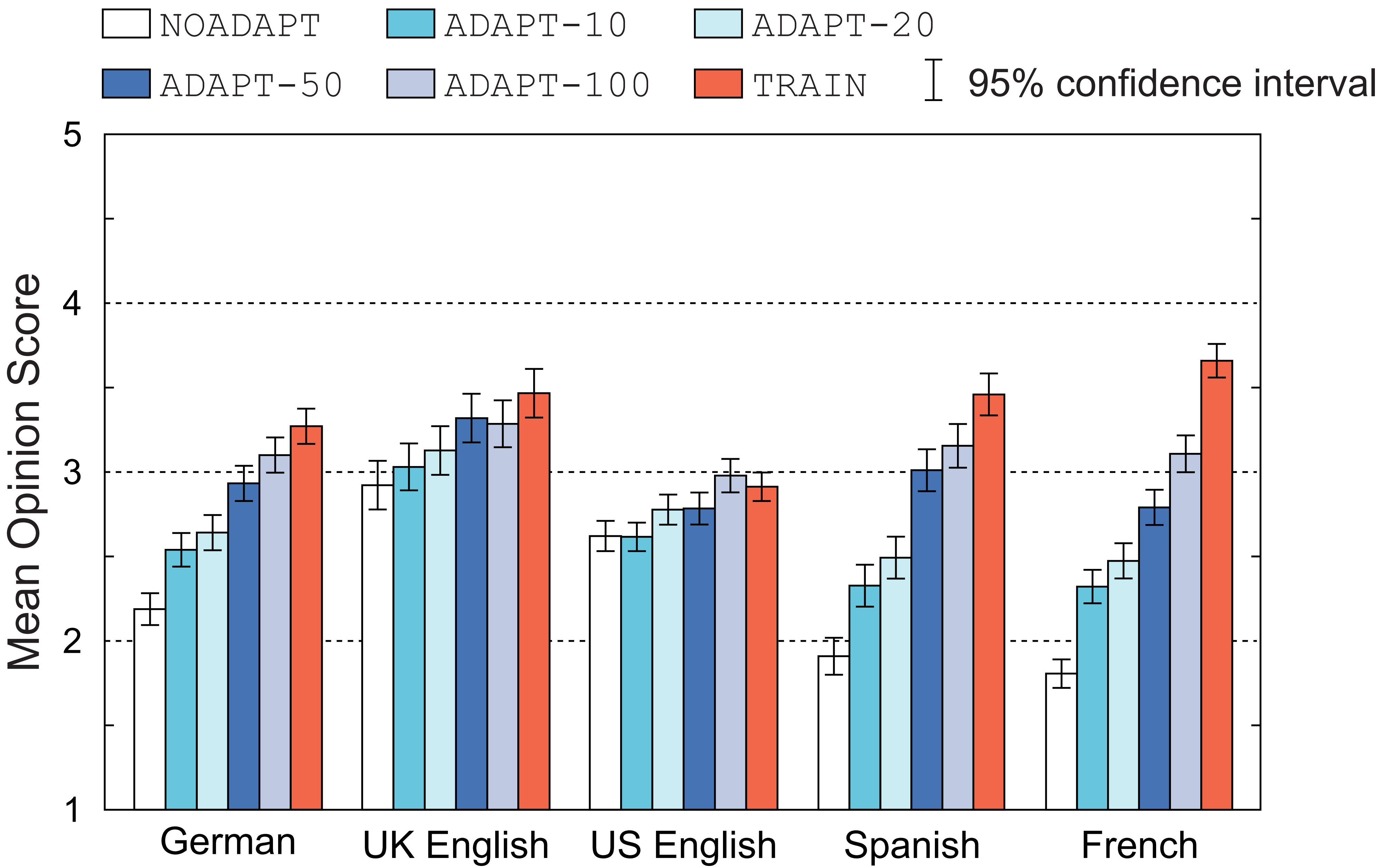
Language	Parameter	clusters 2	3	4	5
UK English (training)	mel-cep.	**.597**	.424	.265	.242
	log F0	**.887**	.178	.039	.095
US English (training)	mel-cep.	.468	**.800**	.255	.258
	log F0	.207	**.867**	.049	.100
Spanish (training)	mel-cep.	.332	.148	**.672**	.366
	log F0	.099	.113	**.946**	.078
French (training)	mel-cep.	.244	.001	.403	**.756**
	log F0	.081	.142	.067	**.936**
German (target)	mel-cep.	.352	.271	.308	**.356**
	log F0	.076	**.133**	.028	.063

Preference Test

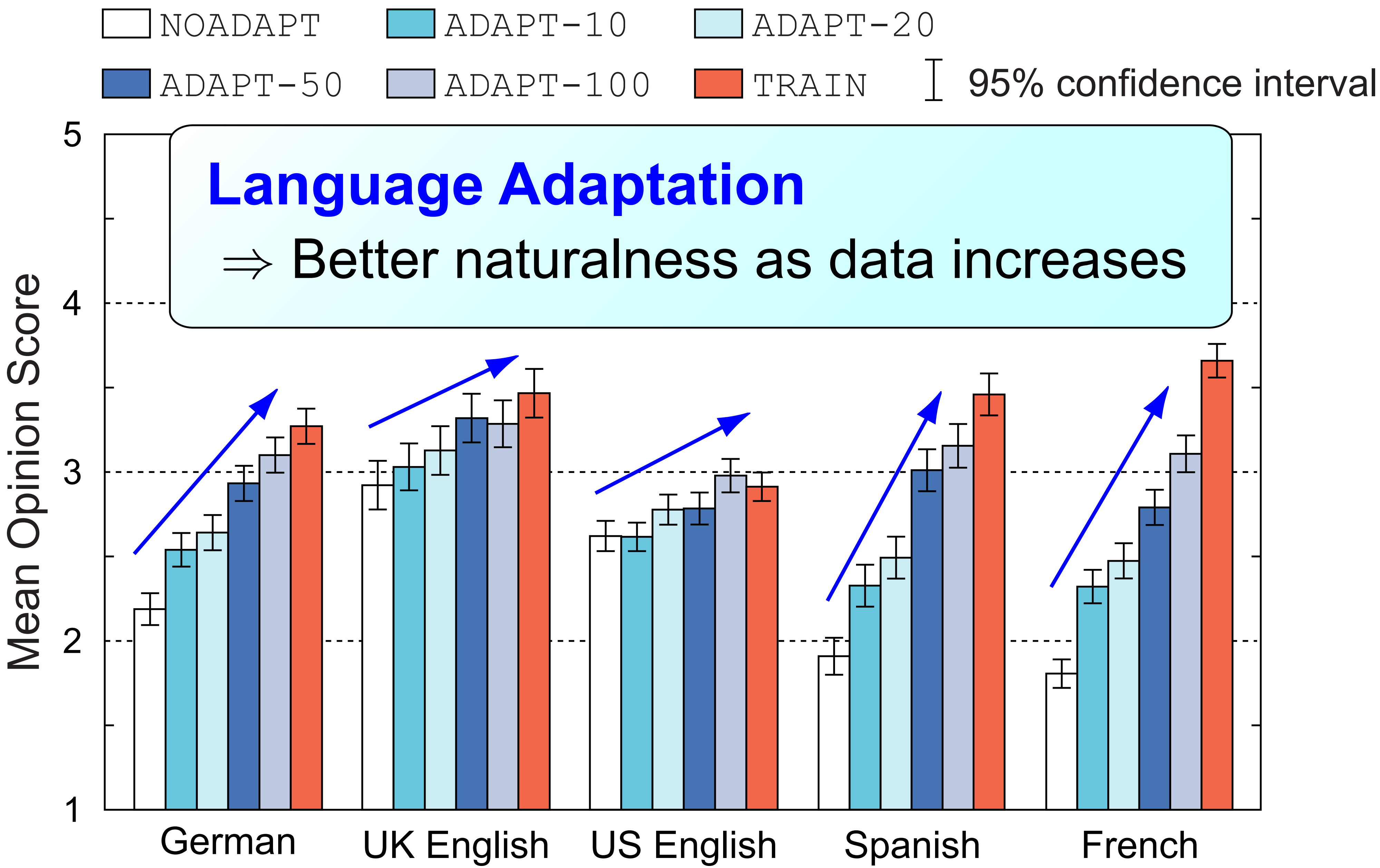
Adaptation data (x 8 spkrs)	Weights only	Weights + Tree	No pref.	*p* (*t*-test)
US En. 10 utts.	36.4	**45.1**	18.5	0.003
US En. 50 utts.	32.8	**51.8**	15.4	< 0.001
German 10 utts.	25.9	**51.1**	23.0	< 0.001
German 50 utts.	22.4	**69.9**	7.4	< 0.001

Building additional tree was effective

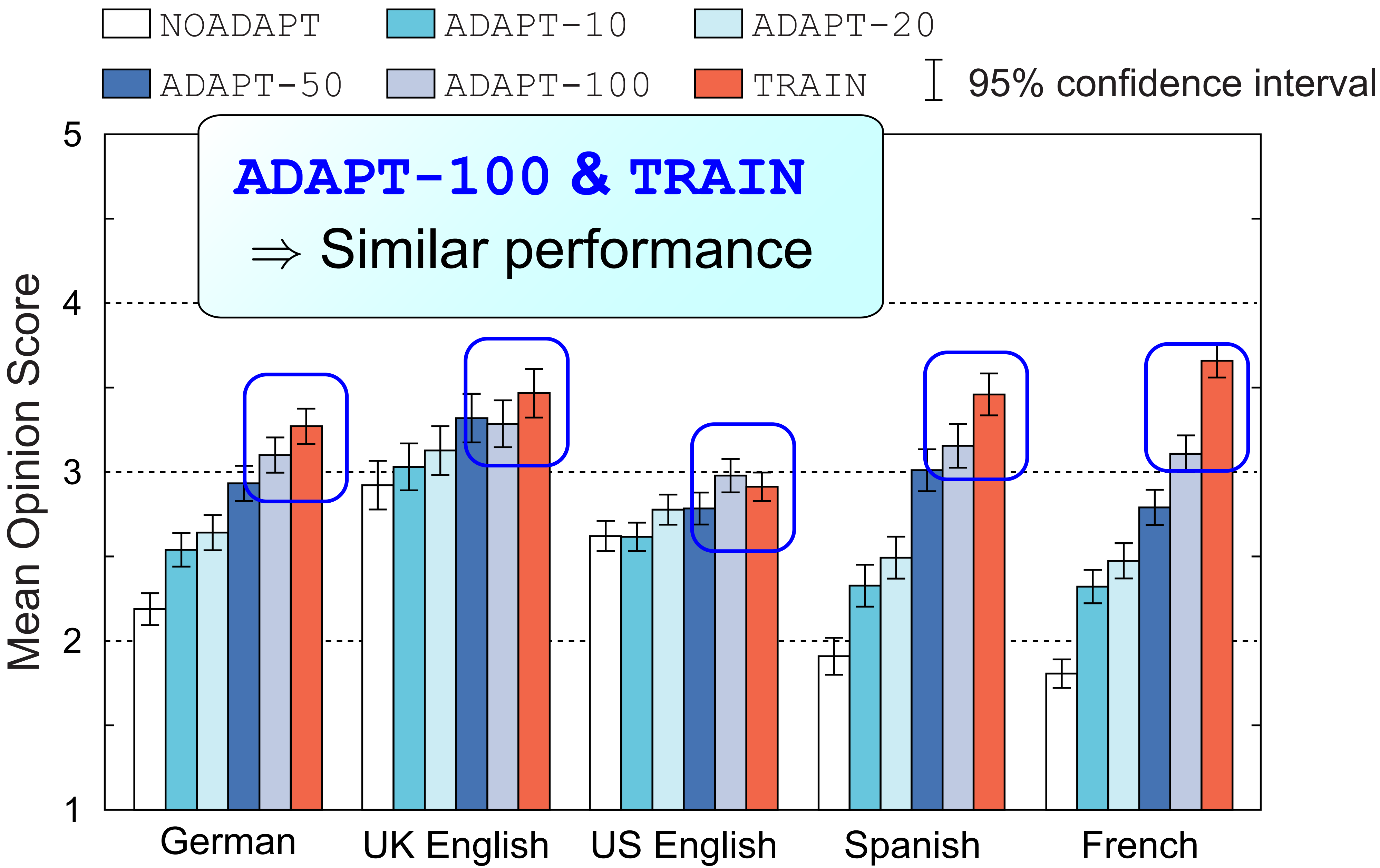
MOS Test


NOADAPT
ADAPT-10
ADAPT-20
ADAPT-50
ADAPT-100
TRAIN
95% confidence interval
Mean Opinion Score
5
4
3
2
1
German
UK English
US English
Spanish
French

MOS Test


NOADAPT
ADAPT-10
ADAPT-20
ADAPT-50
ADAPT-100
TRAIN
95% confidence interval
Language Adaptation
⇒ Better naturalness as data increases
Mean Opinion Score
5
4
3
2
1
German
UK English
US English
Spanish
French

MOS Test



ADAPT-100 & TRAIN

$\Rightarrow$ Similar performance

Outline

- Background
- Conventional approaches
    * Polyglot speaker
    * Mixing mono-lingual corpora
    * Cross-lingual speaker adaptation
- Speaker & language factorization (SLF)
    * Concept
    * Details
- Experiments
- Conclusions

Conclusions

Speaker & language factorization (SLF)

- Application of acoustic factorization to speech synthesis
- Combine 2 transforms
  * CMLLR based speaker transform
  * CAT w/ cluster-dependent trees for language transform
- Better naturalness by increasing amount of data
- Polyglot synthesis
- Adaptation to new languages

Future plans

- Increase amount of data & # of speakers per language
- Add more languages (e.g., Japanese, Mandarin)

References

[Anastasakos;'96] T. Anastasakos, et al., "A compact model for speaker adaptive training," in Proc. ICSLP, 1996, pp. 1137-1140.
[Gales;'01] M. Gales, "Acoustic factorisation," in Proc. ASRU, 2001, pp. 77-80.
[Traber;'99] C. Traber, et al., "From multilingual to polyglot speech synthesis," in Proc. Eurospeech, 1999, pp. 835-838.
[Latorre;'06] J. Latorre, et al., "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer," Speech Communication, vol. 48, no. 10, pp. 1227-1242, 2006.
[Black;'06] A. Black, et al., "Speaker clustering for mulitilingual synthesis," in Proc. ISCA ITRW MULTILING, 2008.
[Chen;'09] Y.-N. Chen, et al. "State mapping for cross-language speaker adaptation in TTS," in Proc. ICASSP, 2009, pp. 4273-4276.
[Wu;'09] Y.-J. Wu, et al., "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in Proc. Interspeech, 2009, pp. 528-531.
[Gales;'98] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Computer Speech & Language, vol. 12, no. 2, pp. 75-98, 1998.
[Zen;'09] H. Zen, et al., "Context-dependent additive log F0 model for HMM-based speech synthesis," in Proc. Interspeech, 2009, pp. 2091-2094.
[Gales;'00] M. Gales, "Cluster adaptive training of hidden Markov models," IEEE Trans. Speech Audio Processing, vol. 8, no. 4, pp. 417-428, 2000.