Multilingual and low-resource ASR

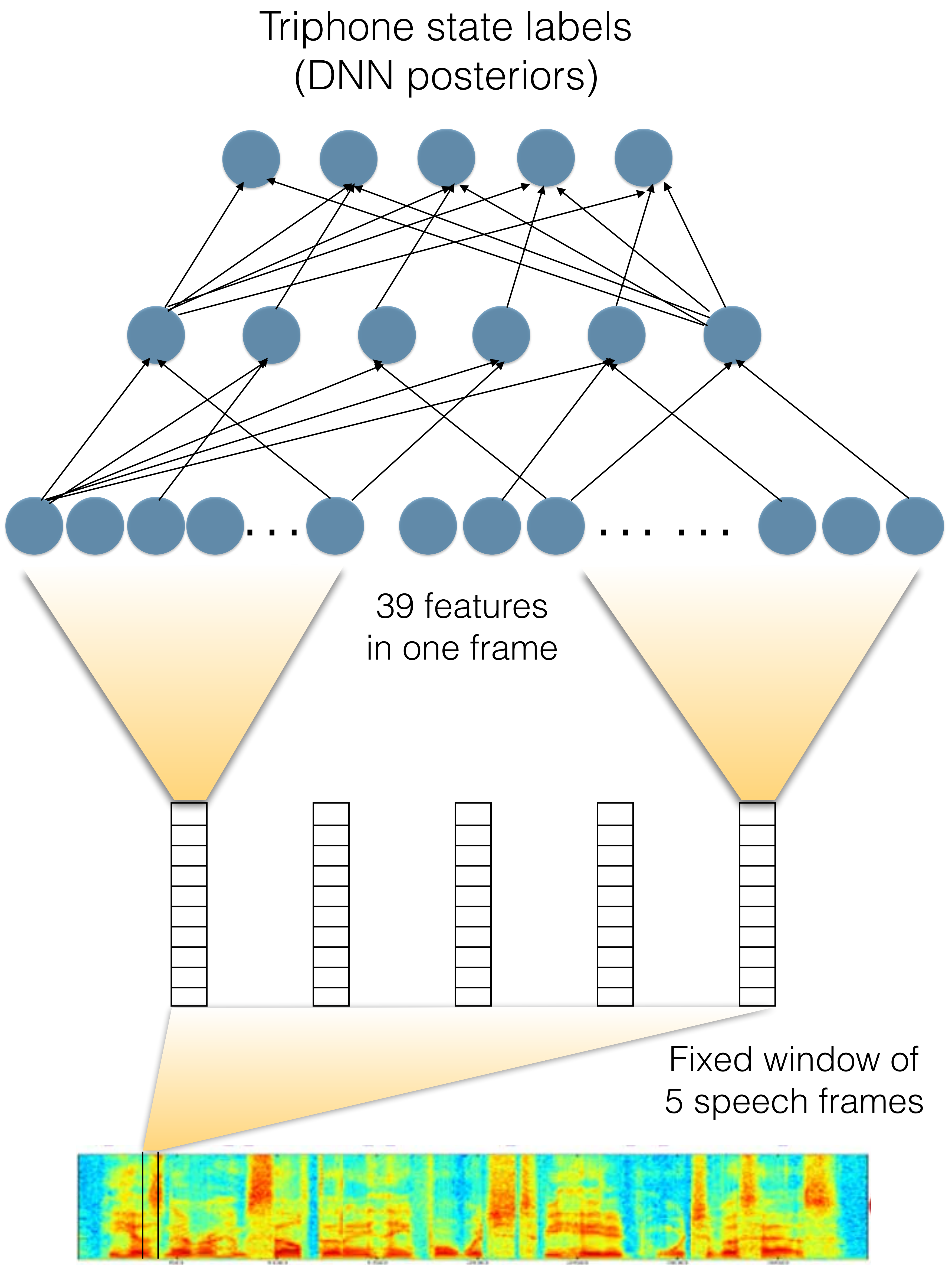
Lecture 18

CS 753

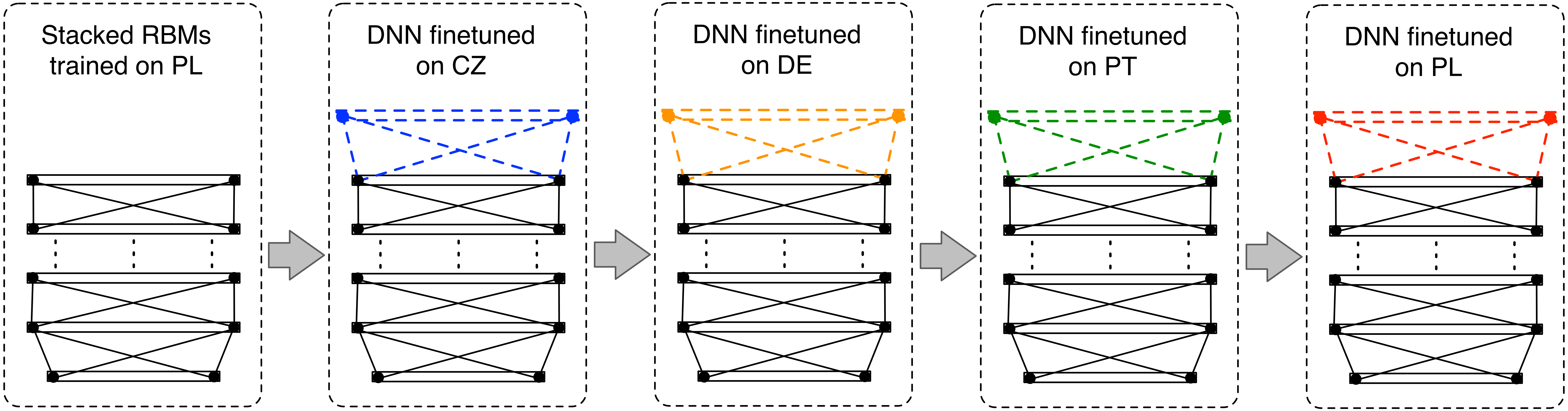
Instructor: Preethi Jyothi

Recall Hybrid DNN-HMM Systems

- Instead of GMMs, use scaled DNN posteriors as the HMM observation probabilities
- DNN trained using triphone labels derived from a forced alignment “Viterbi” step.

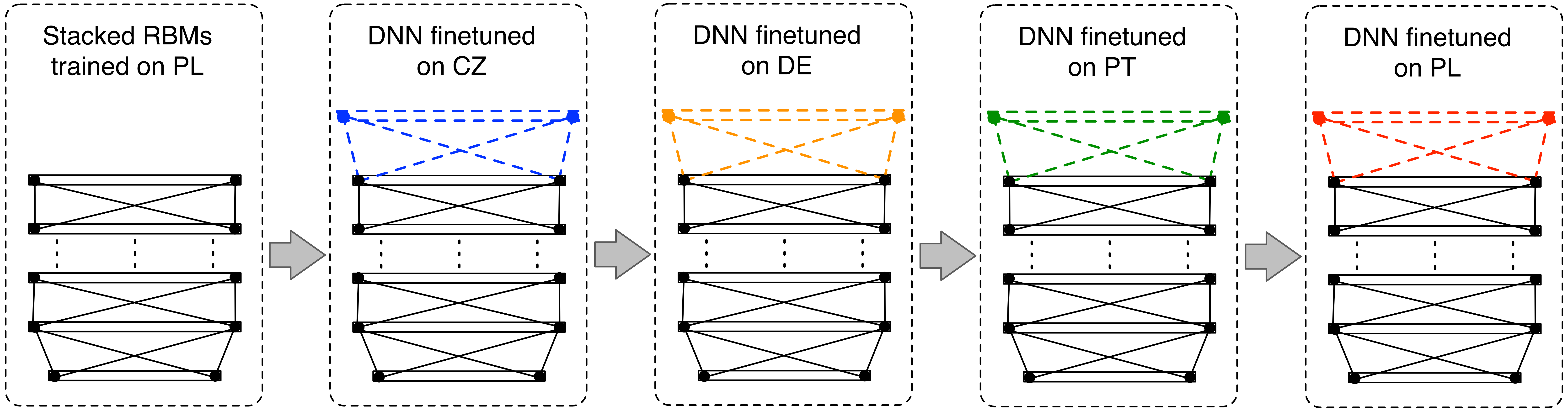


Multilingual Training
(Hybrid DNN/HMM System)



Image/Table from Ghoshal et al., “Multilingual training of deep neural networks”, ICASSP, 2013.

Multilingual Training (Hybrid DNN/HMM System)



Different training language schedules

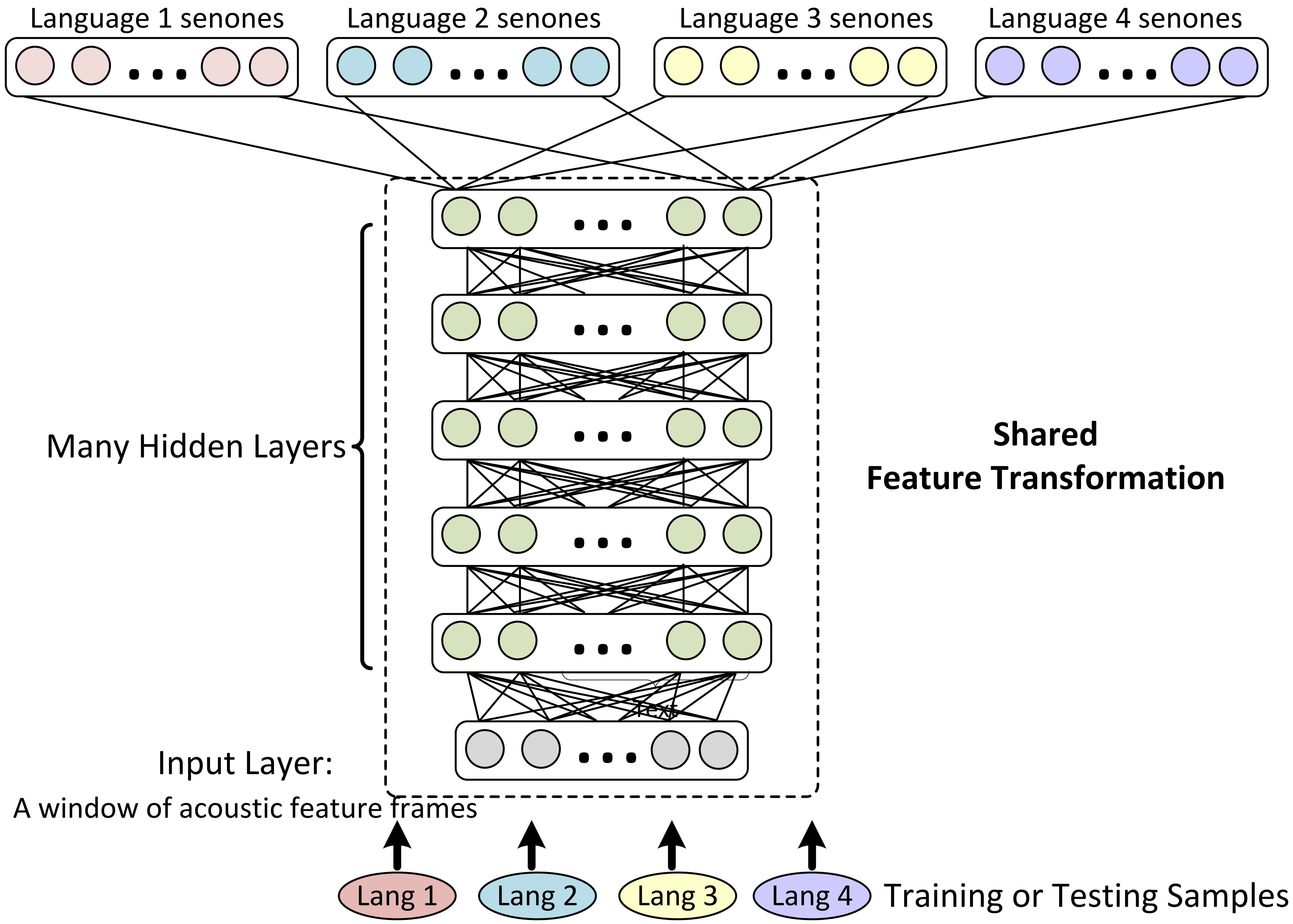
Languages	Dev	Eval
RU	27.5	24.3
CZ →RU	27.5	24.6
CZ →DE →FR →SP →RU	26.6	23.8
CZ →DE →FR →SP →PT →RU	26.3	23.6

Mono- and multilingual results

Language	Vocab	PPL	ML-GMM WER(%)	DNN WER(%)	Multilingual DNN Languages	Multilingual DNN WER(%)
CZ	29K	823	18.5	15.8	—	—
DE	36K	115	13.9	11.2	CZ →DE	9.4
FR	16K	341	25.8	22.6	CZ →DE →FR	22.6
SP	17K	134	26.3	22.3	CZ →DE →FR →SP	21.2
PT	52K	184	24.1	19.1	CZ →DE →FR →SP →PT	18.9
RU	24K	634	32.5	27.5	CZ →DE →FR →SP →PT →RU	26.3
PL	29K	705	20.0	17.4	CZ →DE →FR →SP →PT →RU →PL	15.9

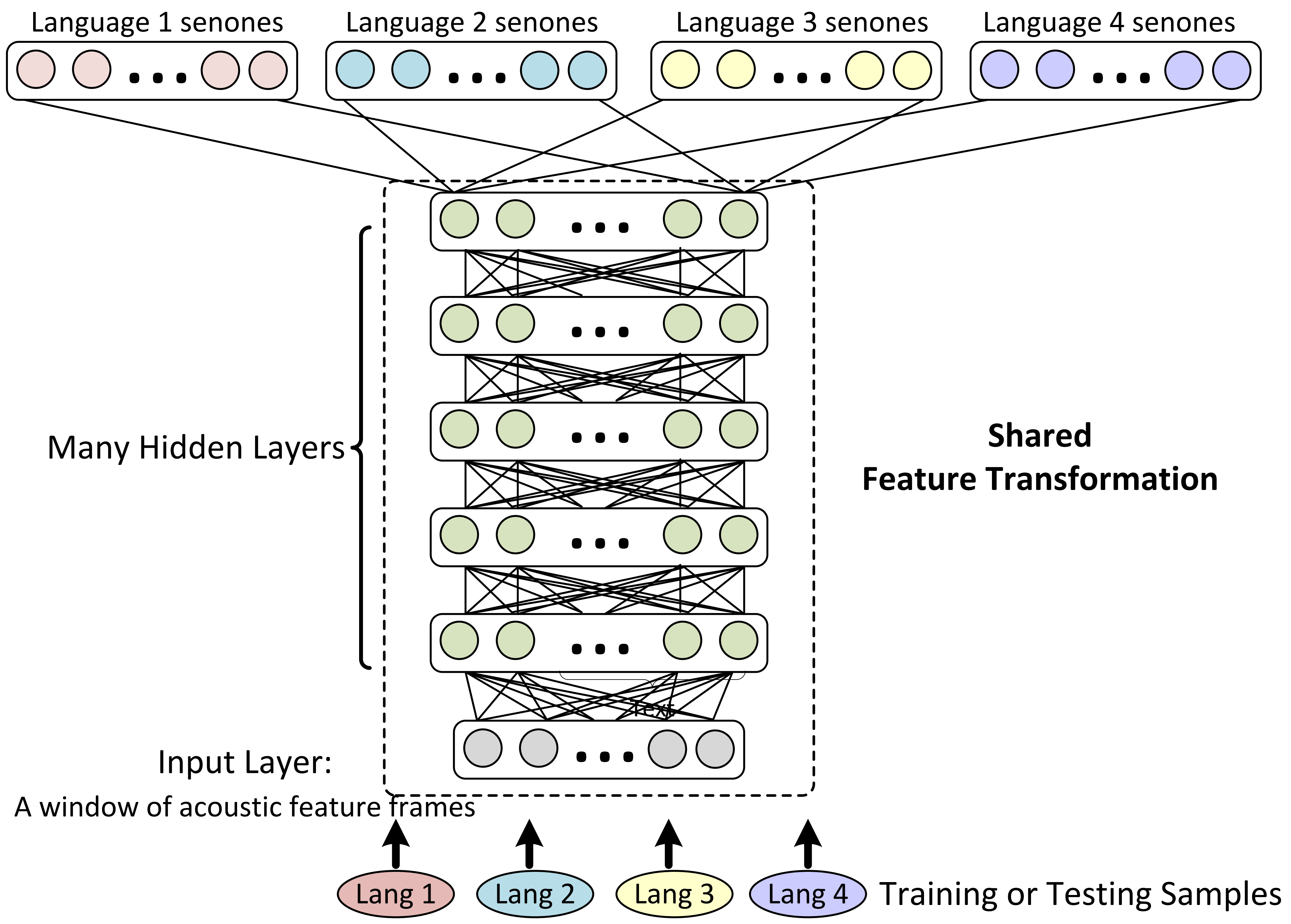
Image/Table from Ghoshal et al., "Multilingual training of deep neural networks", ICASSP, 2013.

Shared hidden layers + Language-specific softmax layers



Huang et al., “Cross-language knowledge transfer using multilingual DNNs with shared hidden layers”, ICASSP 2013.

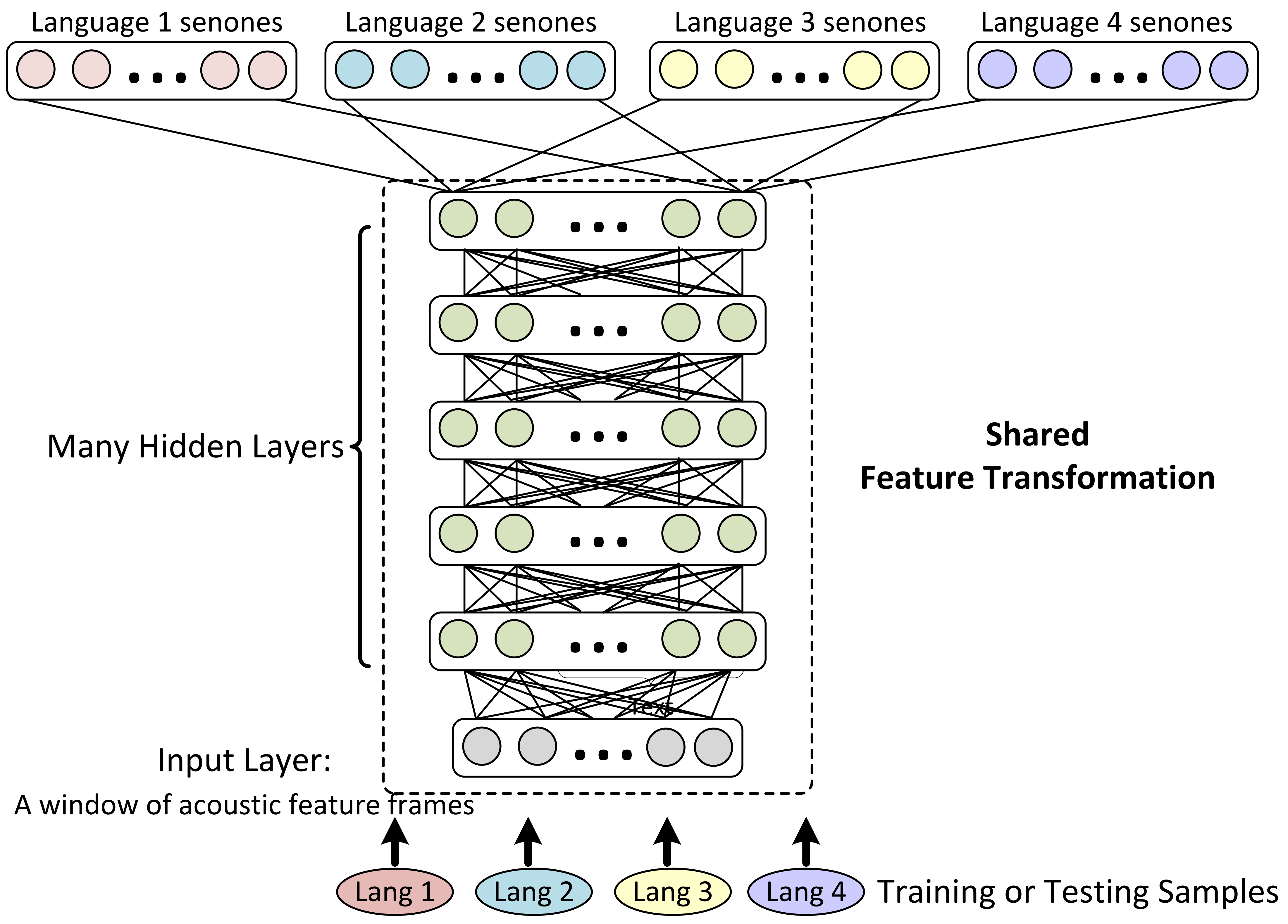
Shared hidden layers + Language-specific softmax layers



- Hidden layers are shared across languages; treated as a universal feature transformation
- Each language has its own softmax layer to estimate posterior probabilities of tied triphone states specific to each language

Huang et al., "Cross-language knowledge transfer using multilingual DNNs with shared hidden layers", ICASSP 2013.

Shared hidden layers + Language-specific softmax layers



Hidden layers are transferable

	WER (%)
Baseline (9-hr ENU)	30.9
FRA HLs + Train All Layers	30.6
FRA HLs + Train Softmax Layer	27.3
SHL-MDNN + Train Softmax Layer	25.3

Training strategy based on target language data

ENU training data (#. Hours)	3	9	36
Baseline DNN (no Transfer)	38.9	30.9	23.0
SHL-MDNN + Train Softmax Layer	**28.0**	**25.3**	22.4
SHL-MDNN + Train All Layers	33.4	28.9	**21.6**
Best Case Relative WER Reduction (%)	28.0	18.1	6.1

Cross-lingual transfer

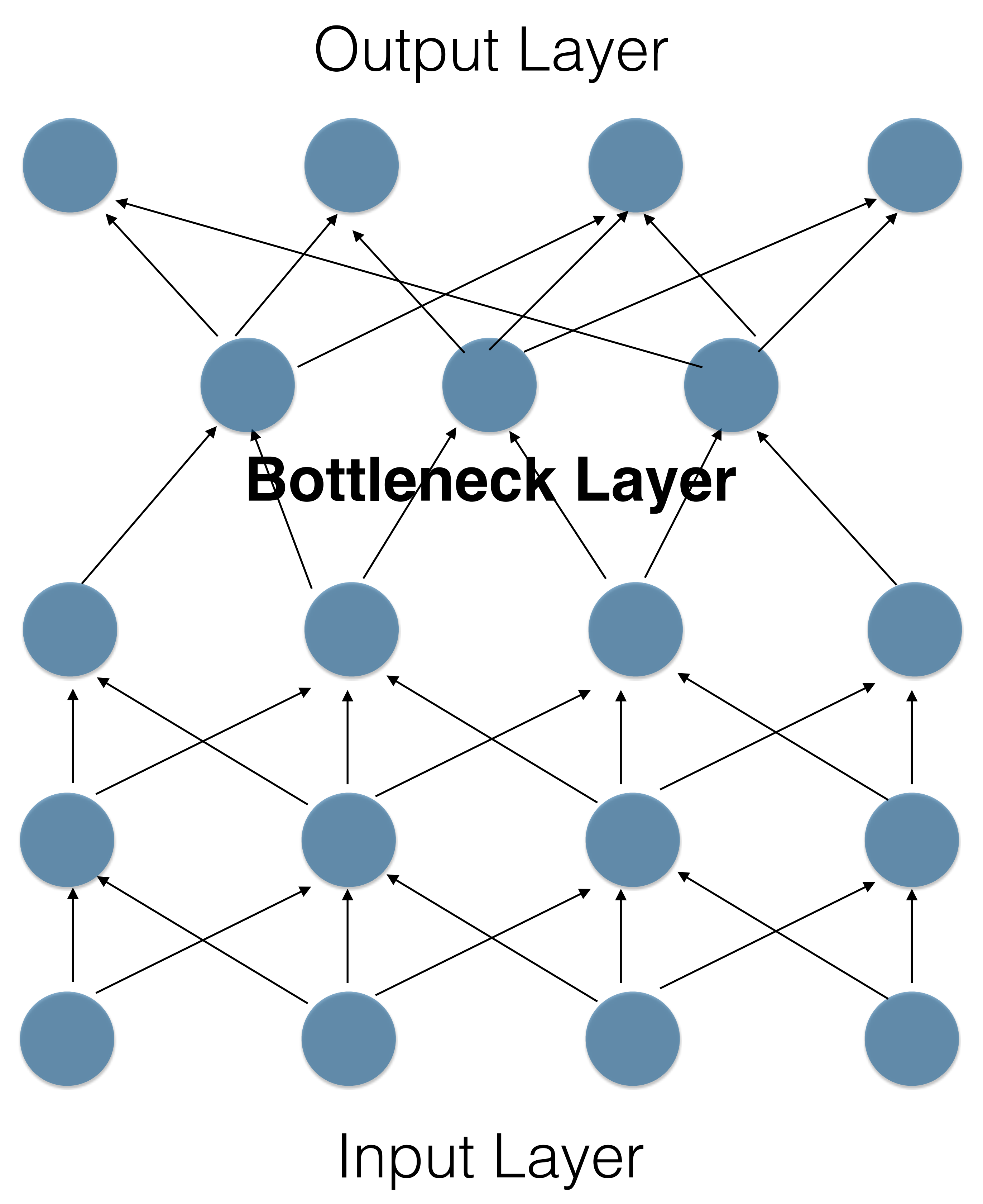
CHN Training Set (Hrs)	3	9	36	139
Baseline - CHN only	45.1	40.3	31.7	29.0
SHL-MDNN Model Transfer	35.6	33.9	28.4	26.6
Relative CER Reduction	21.1	15.9	10.4	8.3

Huang et al., "Cross-language knowledge transfer using multilingual DNNs with shared hidden layers", ICASSP 2013.

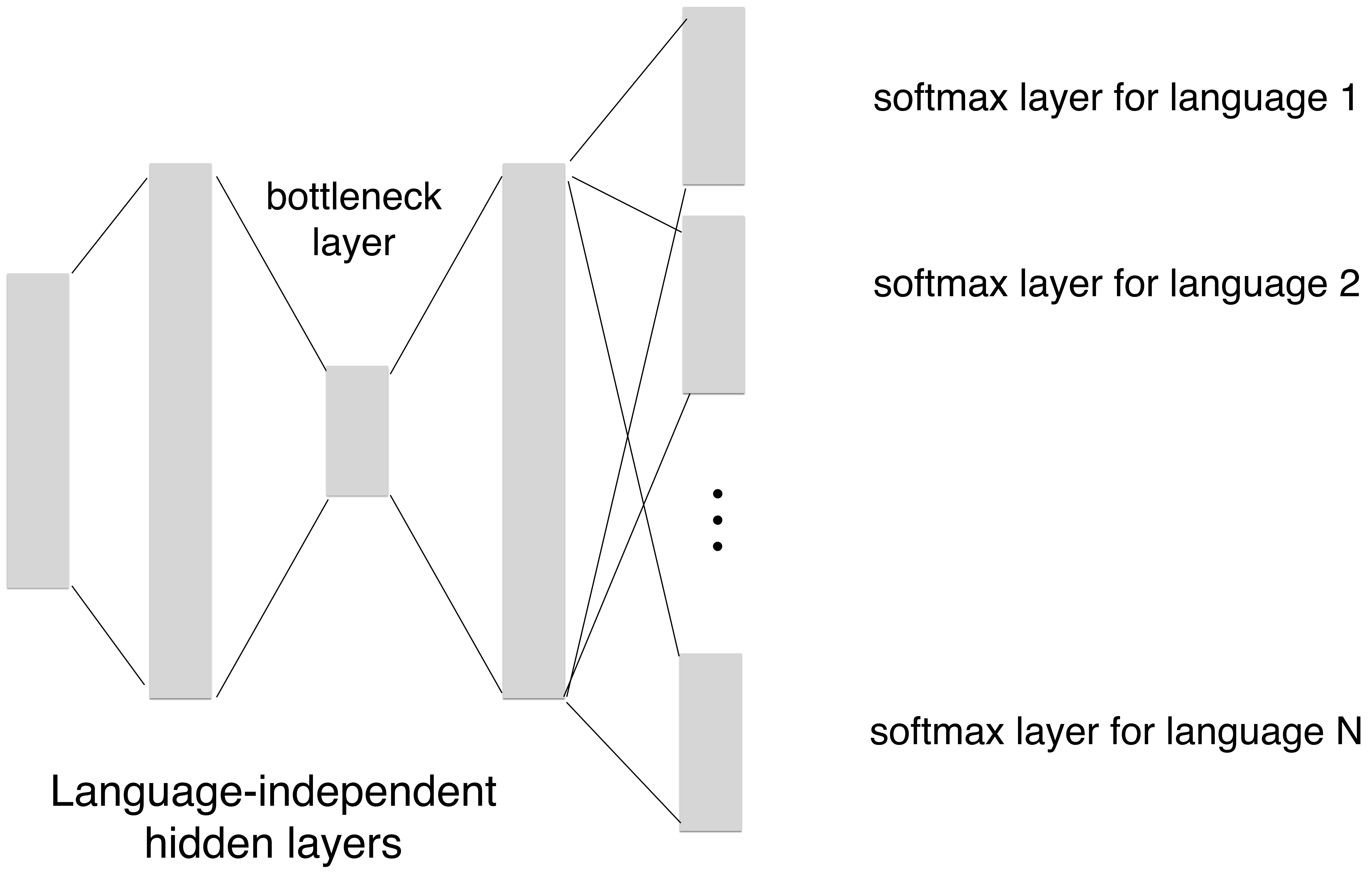
Recall Tandem DNN-HMM Systems

- Neural network outputs are used as "features" to train HMM-GMM models

- Use a low-dimensional bottleneck layer representation to extract features from the bottleneck layer

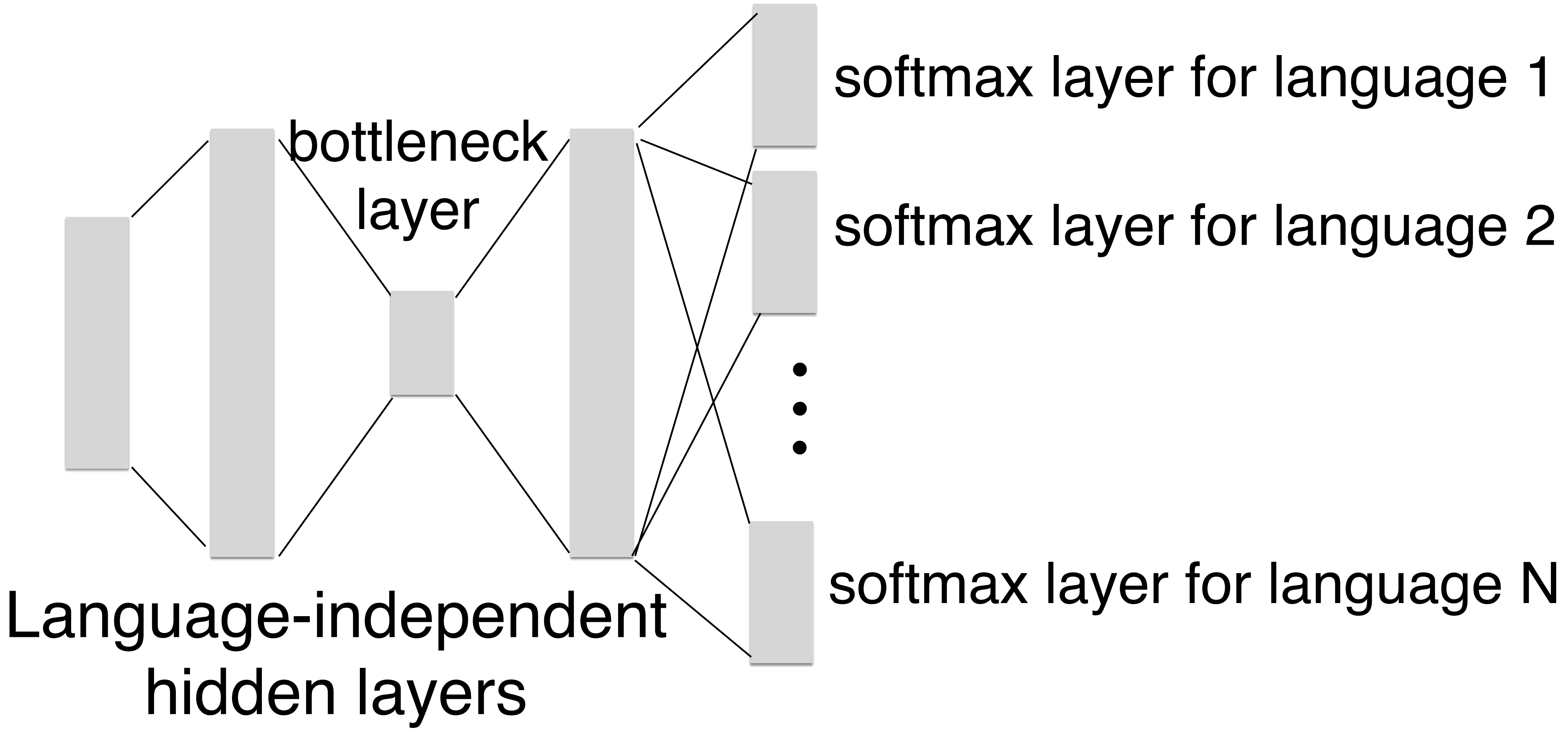


Multilingual Training (Tandem System)



Vesely et al., “The language-independent bottleneck features”, SLT, 2012.

Multilingual Training (Tandem System)

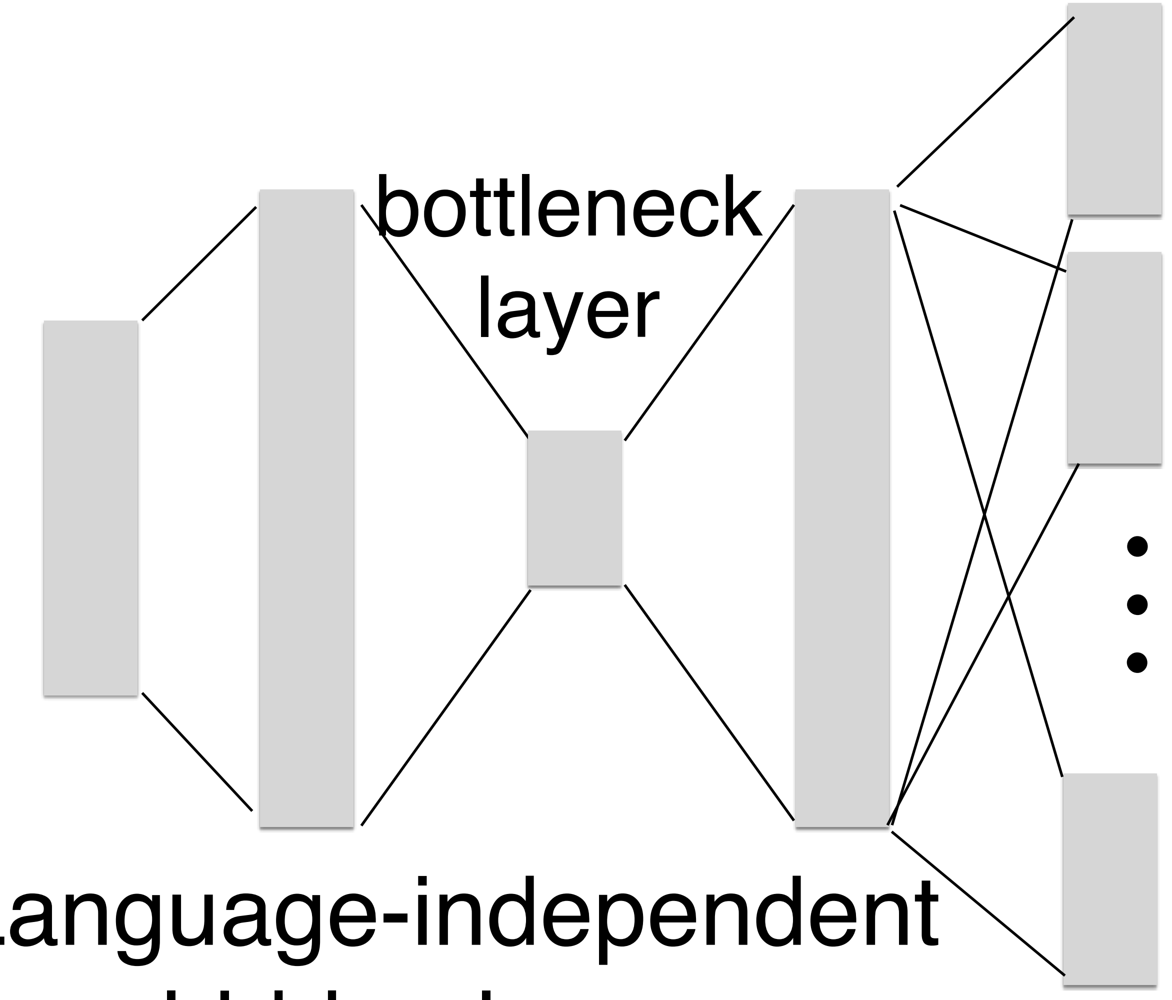


Monolingual/multilingual BN feature-based results

Language	Czech	English	German	Portugese	Spanish	Russian	Turkish	Vietnamese
HMM	22.6	16.8	26.6	27.0	23.0	33.5	32.0	27.3
1-Softmax	20.3	16.1	25.9	27.2	24.2	33.4	31.3	26.9
mono-BN	19.7	15.9	25.5	27.2	23.2	32.5	30.4	23.4
1-Softmax(IPA)	19.4	15.5	24.8	25.6	23.2	32.5	30.3	25.9
8-Softmax	**19.3**	**14.7**	**24.0**	**25.2**	**22.6**	**31.5**	**29.4**	**24.3**

Vesely et al., "The language-independent bottleneck features", SLT, 2012.

Multilingual Training (Tandem System)

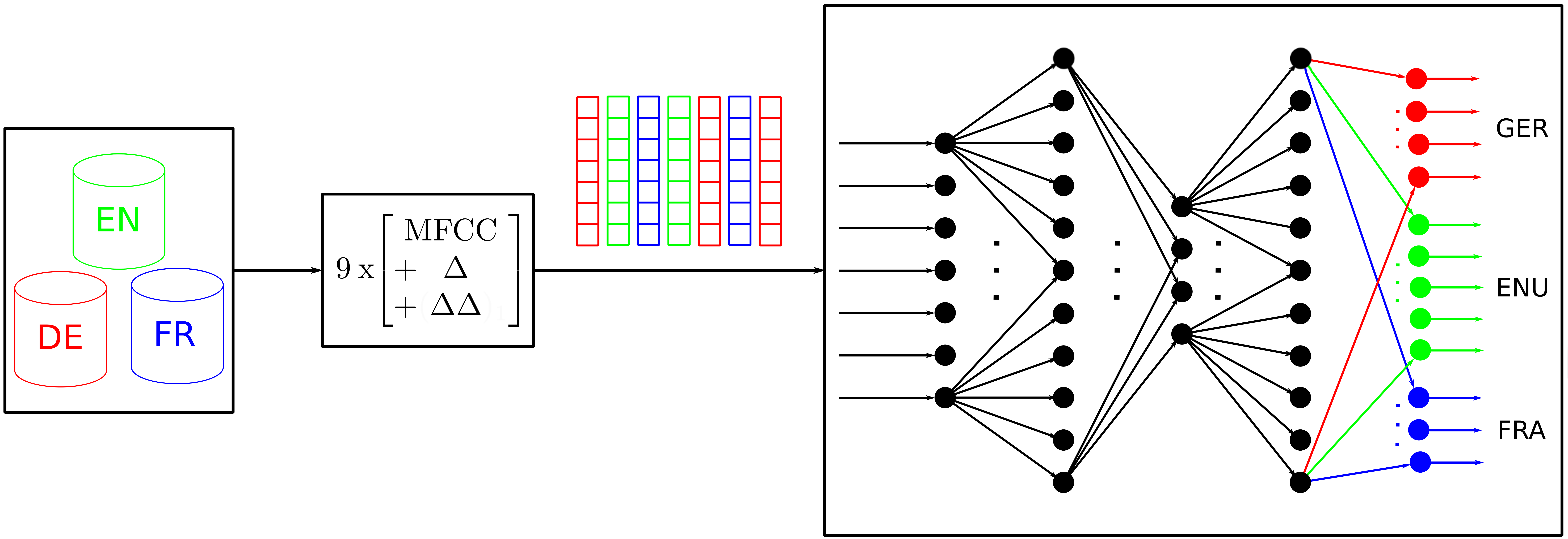


Cross-lingual WERs

Language	baselines: PLP-HLDA (II.)	baselines: Mono-BN (III.)	ANN output : 5-Softmax (lang-pooled) (d)
Czech	22.6	19.7	19.2
English	16.8	15.9	14.7
German	26.6	25.5	24.5
Portuguese	27.0	27.2	26.0
Spanish	23.0	23.2	23.0
Russian	33.5	32.5	**32.3**
Turkish	32.0	**30.4**	30.7
Vietnamese	27.3	**23.4**	26.8

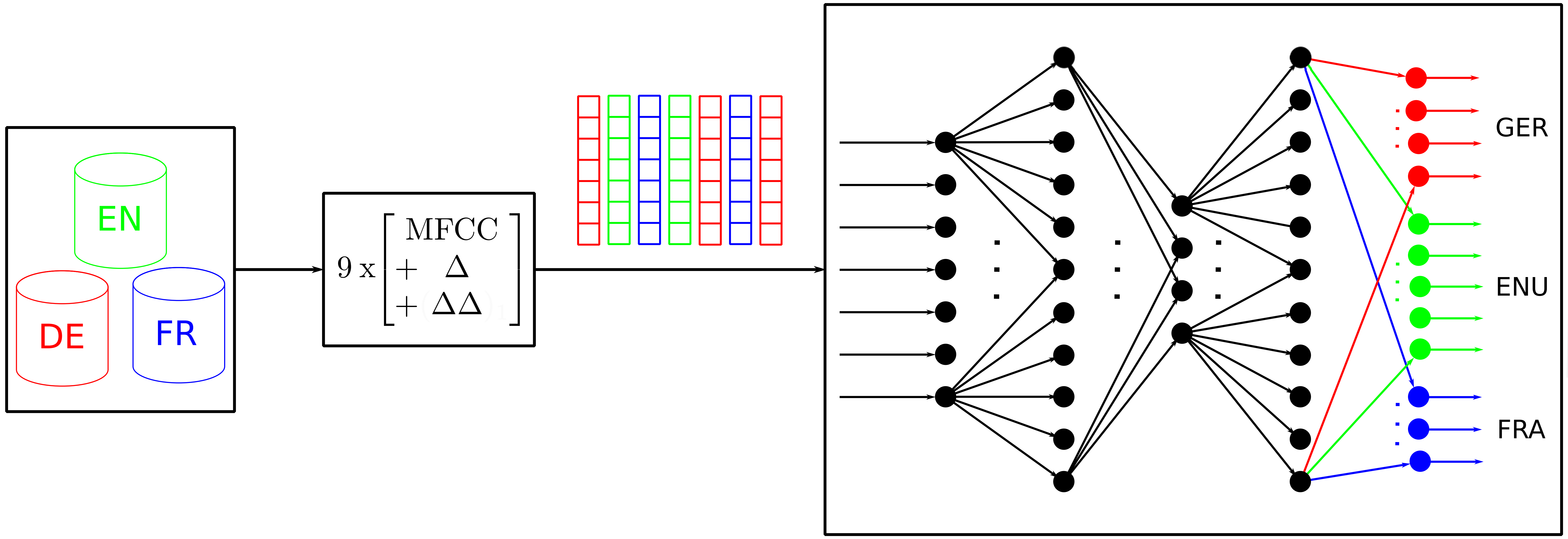
Vesely et al., "The language-independent bottleneck features", SLT, 2012.

Cross- and Multilingual Bottleneck features



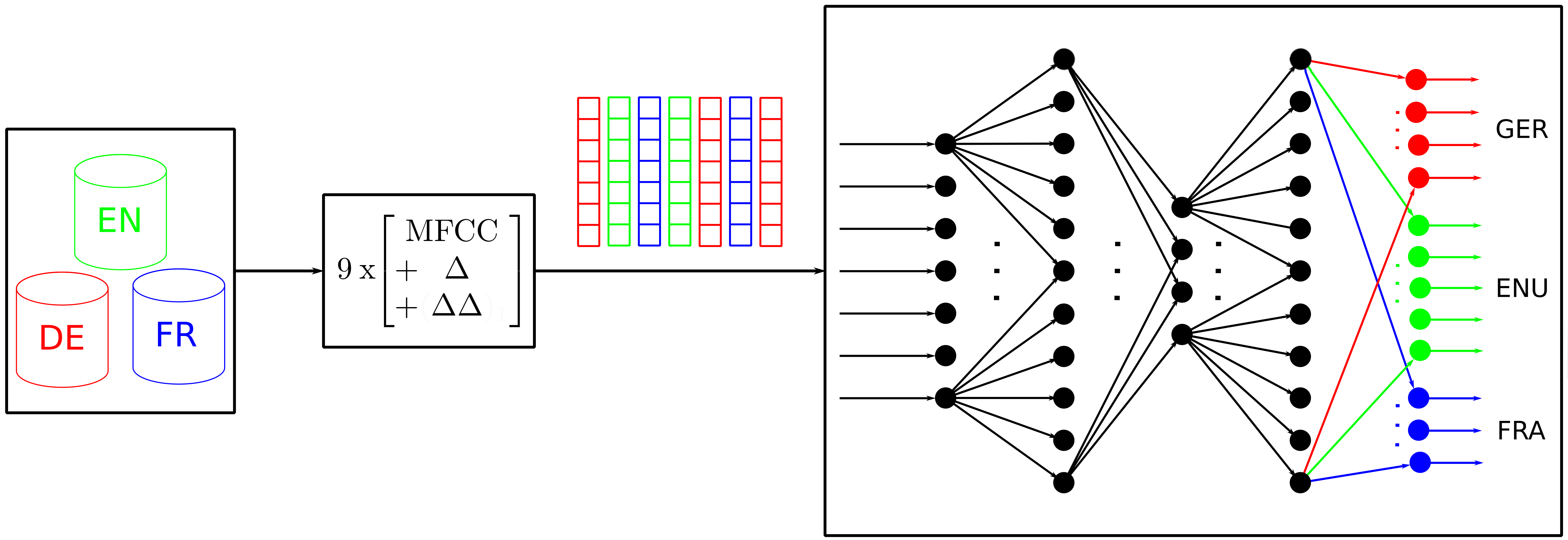
Tuske et al., “Investigation on cross- and multilingual MLP features”, ICASSP, 2013

Cross- and Multilingual Bottleneck features



- Features from three languages are merged and presented as input to the model
- Language-specific softmax layers
- Bottleneck layer which is shared across languages

Cross- and Multilingual Bottleneck features



Target and cross-lingual BN features

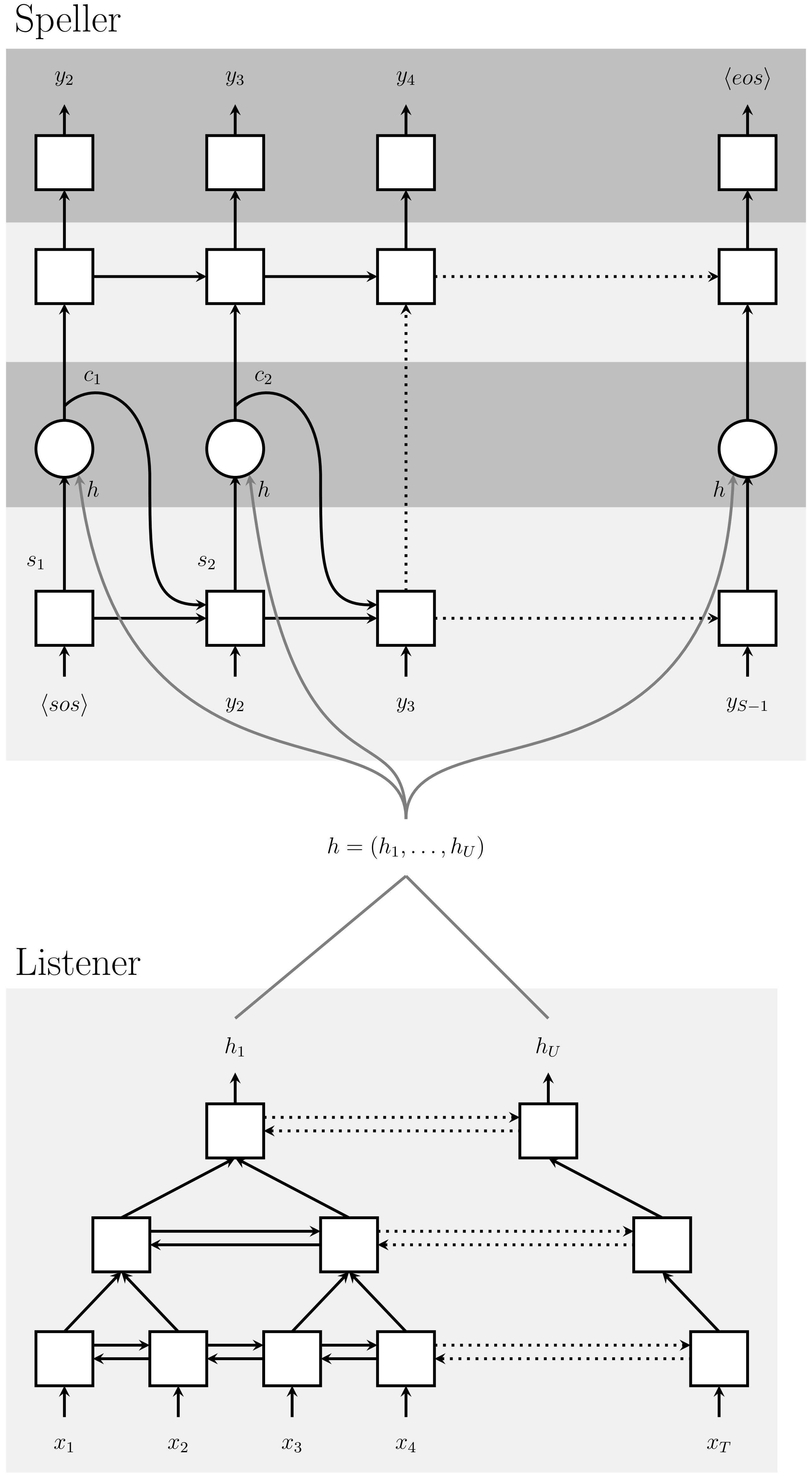
WER [%]		MFCC	MFCC+BN		
			Bottleneck trained on		
			GER	ENU	FRA
Test language	GER	29.97	*27.50* (8.2)	29.63 (1.1)	30.38 (-1.4)
	ENU	21.69	21.31 (1.8)	*18.85* (13.1)	22.63 (-4.3)
	FRA	37.78	37.76 (0.1)	38.72 (-2.5)	*33.95* (10.1)

Multilingual BN features using mismatched data

WER [%]		MFCC+BN			
		BN trained on			
Test language	GER	ENU+FRA	GER+FRA	GER+ENU	GER+ENU+FRA
		28.37 (5.3)	27.06 (9.7)	26.89 (10.3)	26.90 (10.2)
	ENU	GER+FRA	ENU+FRA	ENU+GER	GER+ENU+FRA
		20.29 (6.5)	18.21 (16.0)	17.99 (17.1)	17.89 (17.5)
	FRA	GER+ENU	FRA+GER	FRA+ENU	GER+ENU+FRA
		35.88 (5.0)	33.52 (11.3)	33.45 (11.5)	33.61 (11.0)

e2e multilingual models

Multilingual ASR with an e2e Model

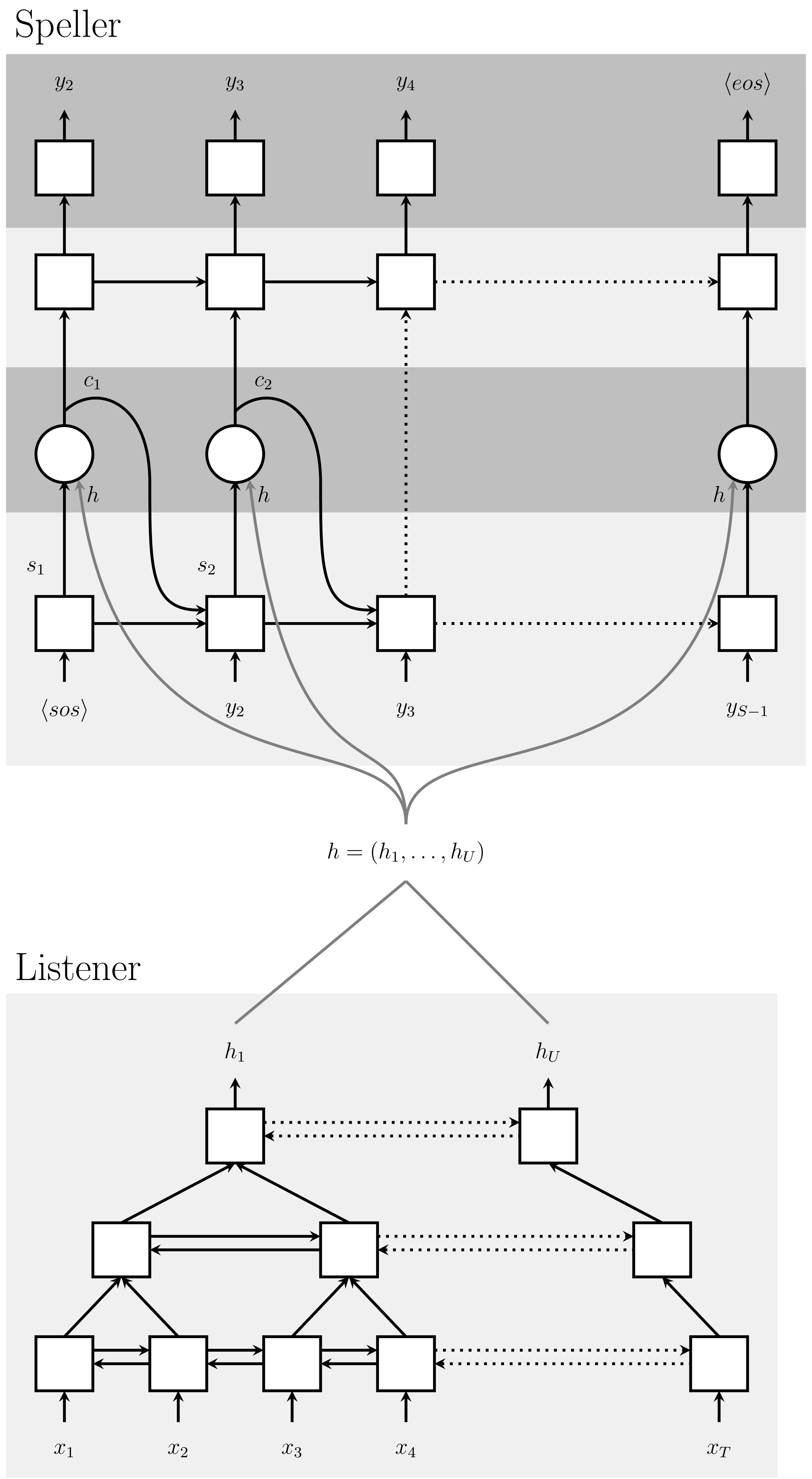


- Use attention-based encoder-decoder models
- Decoder outputs one character per time-step
- For multilingual models, use union over character sets

Bengali	আজ মেঘলা দিন
Gujarati	તે વાદળછાયું દિવસ છે
Hindi	यह एक बादल का दिन है
Kannada	ಇದು ಮೋಡ ಕವಿದ ದಿನ
Malayalam	ഇത് തെളിഞ്ഞ ദിവസമാണ്
Marathi	तो ढगाळ दिवस आहे
Tamil	இது ஒரு மேகமூட்டமான நாள்
Telugu	ఇది మేఘావృతమైన రోజు
Urdu	یہ ابر آلود دن ہے

Image from: Chan et al., Listen, Attend and Spell: A NN for LVCSR, ICASSP 2016

Multilingual ASR with an e2e Model



Language-specific vs. Multilingual models

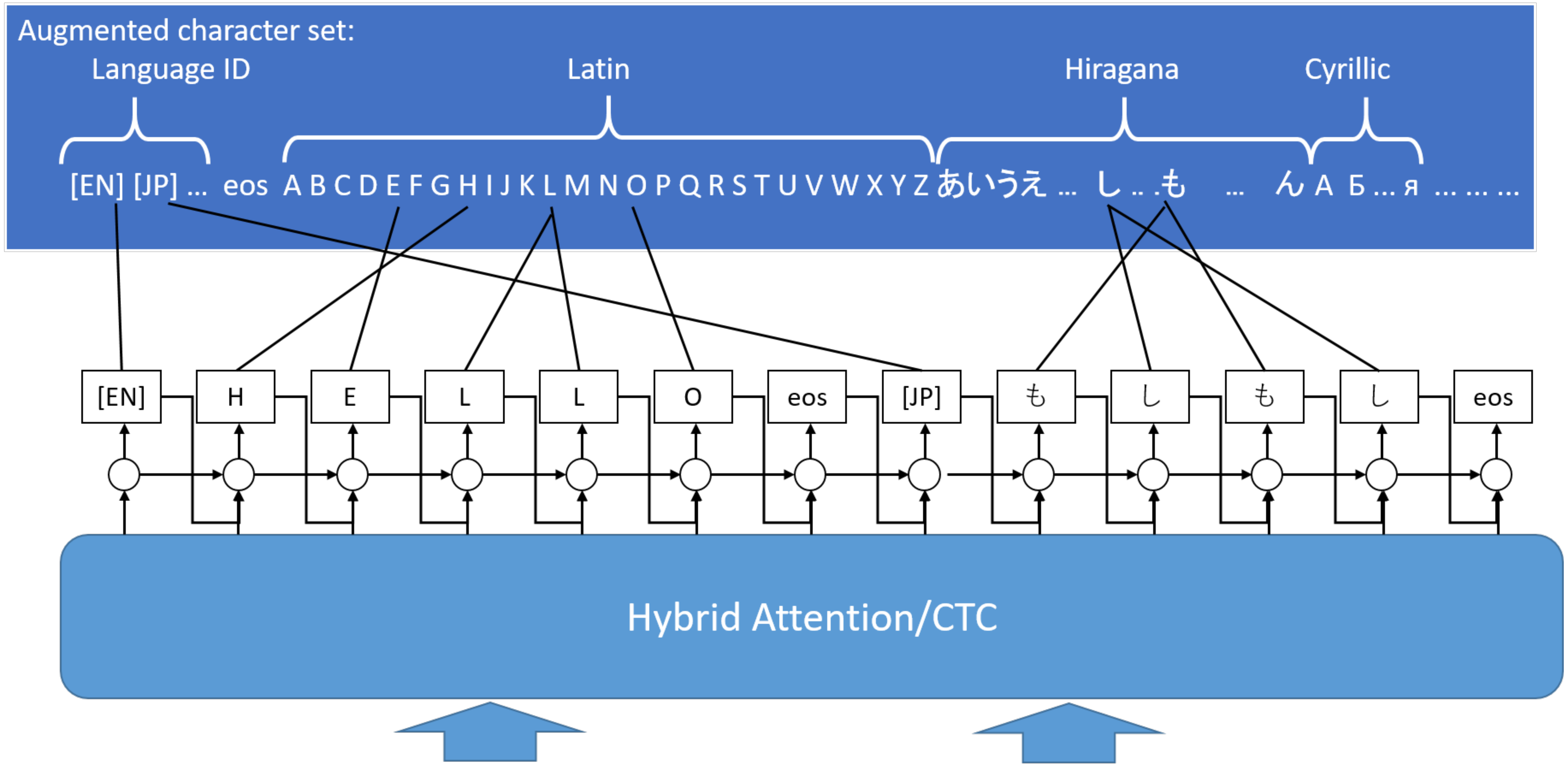
Language	Language-specific	Joint	Joint + MTL
Bengali	19.1	16.8	**16.5**
Gujarati	26.0	**18.0**	18.2
Hindi	16.5	**14.4**	**14.4**
Kannada	35.4	**34.5**	34.6
Malayalam	44.0	36.9	**36.7**
Marathi	28.8	27.6	**27.2**
Tamil	13.3	10.7	**10.6**
Telugu	37.4	**22.5**	22.7
Urdu	29.5	26.8	**26.7**
Weighted Avg.	29.05	**22.93**	**22.91**

LAS models conditioned on language ID

Language	Joint	Dec	Enc	Enc + Dec
Bengali	16.8	16.9	**16.5**	**16.5**
Gujarati	18.0	17.7	**17.2**	17.3
Hindi	**14.4**	14.6	14.5	**14.4**
Kannada	34.5	30.1	29.4	**29.2**
Malayalam	36.9	35.5	34.8	**34.3**
Marathi	27.6	24.0	**22.8**	23.1
Tamil	10.7	10.4	**10.3**	10.4
Telugu	22.5	22.5	21.9	**21.5**
Urdu	26.8	25.7	**24.2**	24.5
Weighted Avg.	22.93	22.03	21.37	**21.32**

Image from: Chan et al., Listen, Attend and Spell: A NN for LVCSR, ICASSP 2016

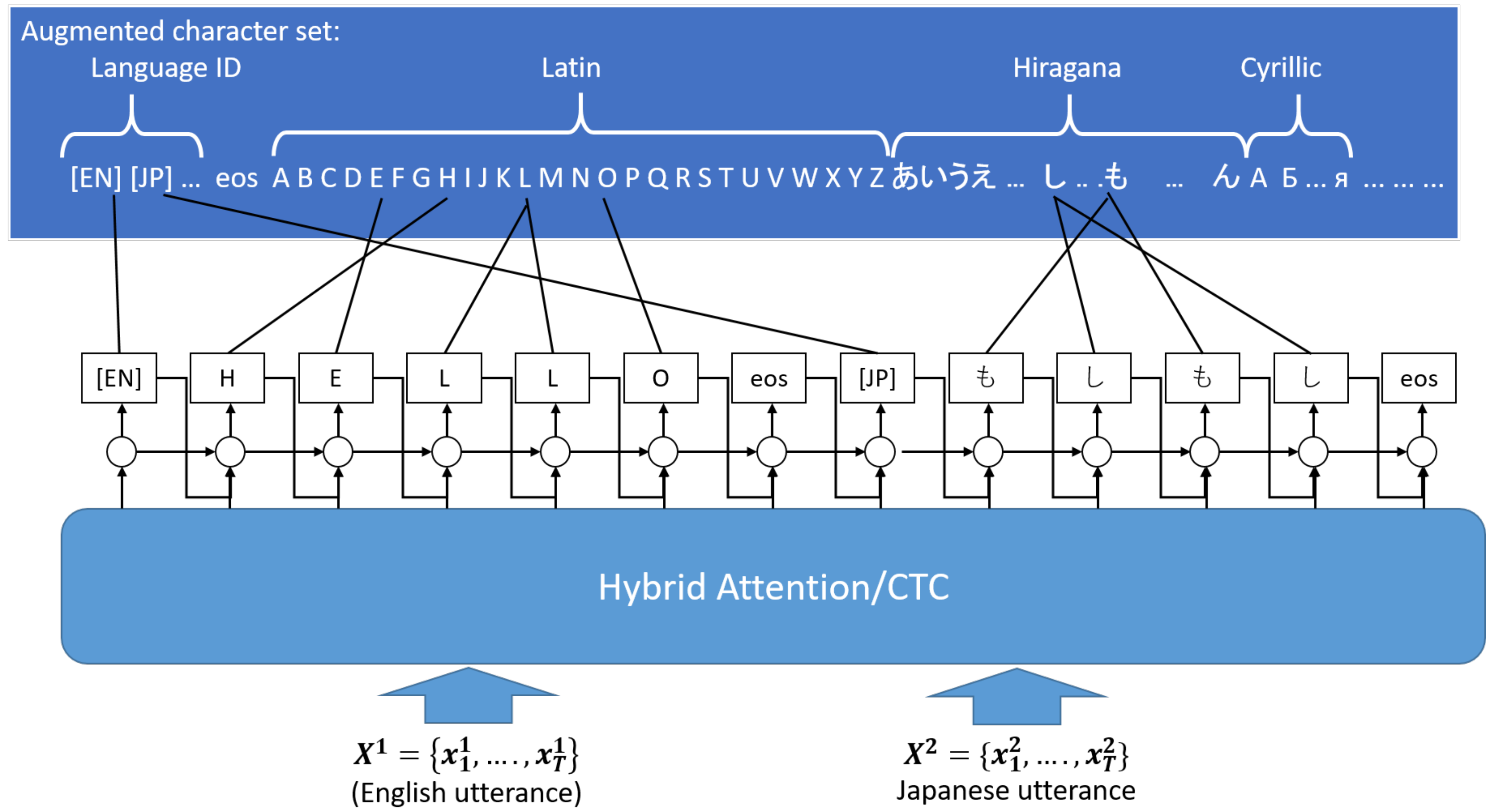
Hybrid End-to-end Multilingual ASR



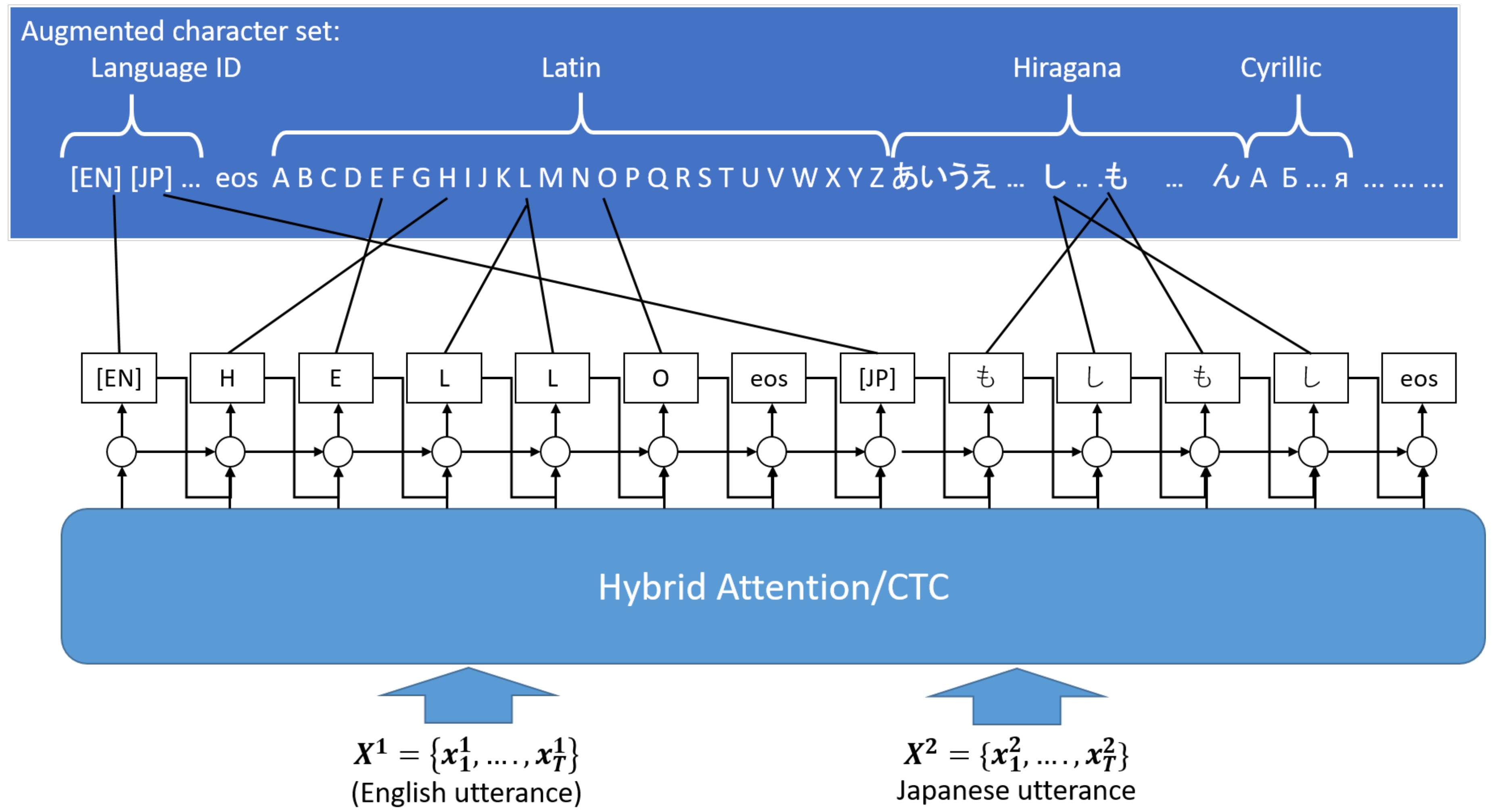
$X^1 = \{x_1^1, \ldots, x_T^1\}$
(English utterance)

$X^2 = \{x_1^2, \ldots, x_T^2\}$
Japanese utterance

Watanabe et al., "e2e architecture for joint language identification and ASR", ASRU, 2017



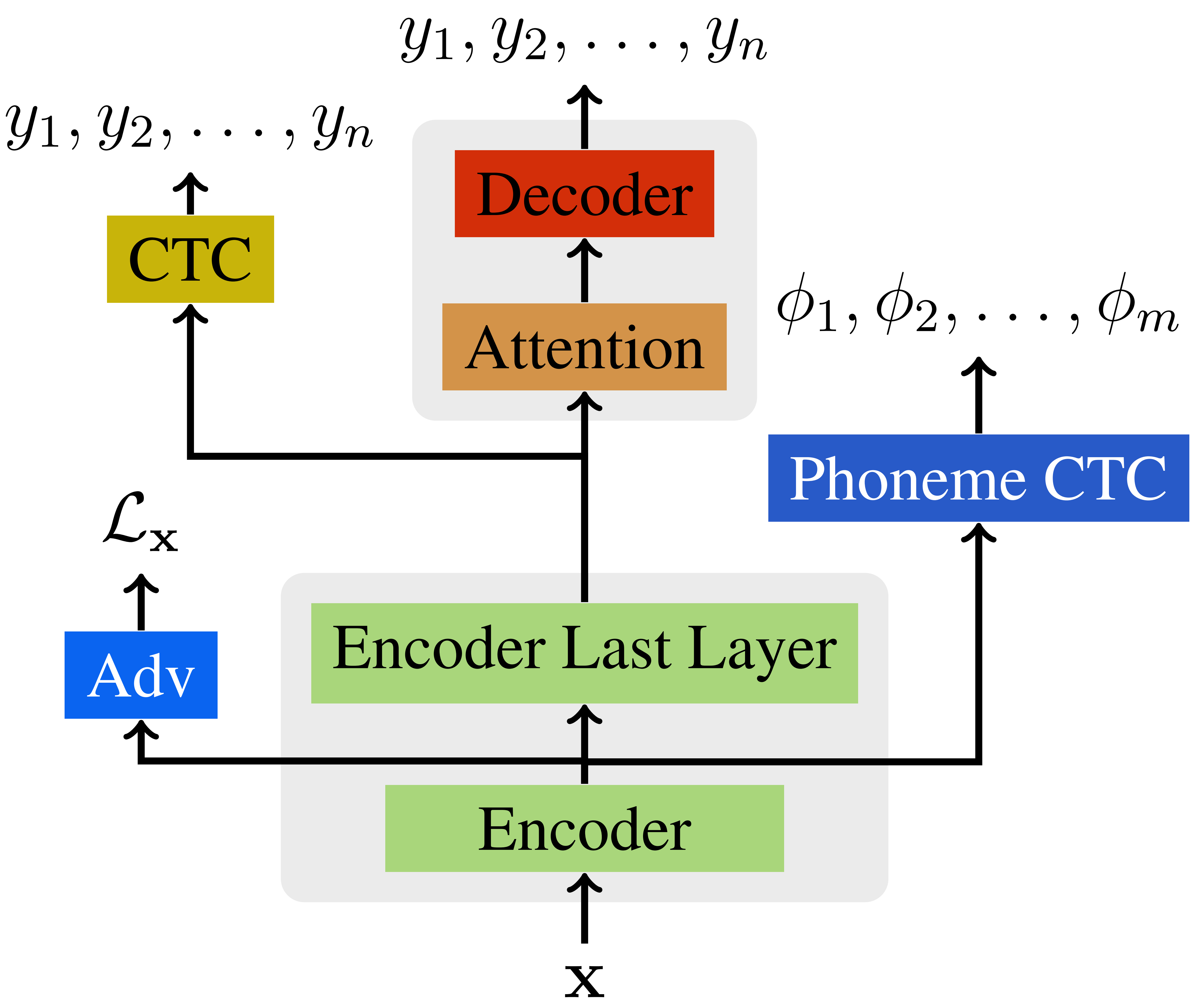
- Hybrid attention+CTC model: Use the CTC objective function as an auxiliary task to train the encoder
- Minimize a linear combination of log-losses of the CTC and attention objectives
- Model also predicts a language ID along with the text outputs



Language-dependent and language-independent CERs

			Language-dependent 4BLSTM	**7lang** 4BLSTM	**7lang** CNN-7BLSTM	**7lang** CNN-7BLSTM RNN-LM	**10lang** CNN-7BLSTM RNN-LM
HKUST	CH	train_dev	40.1	43.9	40.5	40.2	32.0
		dev	40.4	43.6	40.5	40.0	31.0
WSJ	EN	dev93	9.4	9.6	7.7	7.0	9.7
		eval92	7.4	7.3	5.6	5.1	7.4
CSJ	JP	eval1	13.5	14.3	12.4	11.9	10.2
		eval2	10.8	10.8	9.0	8.5	7.2
		eval3	23.2	24.9	22.0	21.4	8.7
Voxforge	DE	dev	6.6	7.4	5.7	5.4	7.3
		eval	5.2	7.4	5.8	5.5	7.3
	ES	dev	50.9	28.1	31.9	31.5	25.8
		eval	50.8	29.6	34.7	34.4	26.7
	FR	dev	27.7	25.0	22.0	21.0	24.1
		eval	26.5	23.5	21.2	20.3	23.2
	IT	dev	14.3	14.3	11.8	11.1	13.8
		eval	14.3	14.4	12.0	11.2	14.1

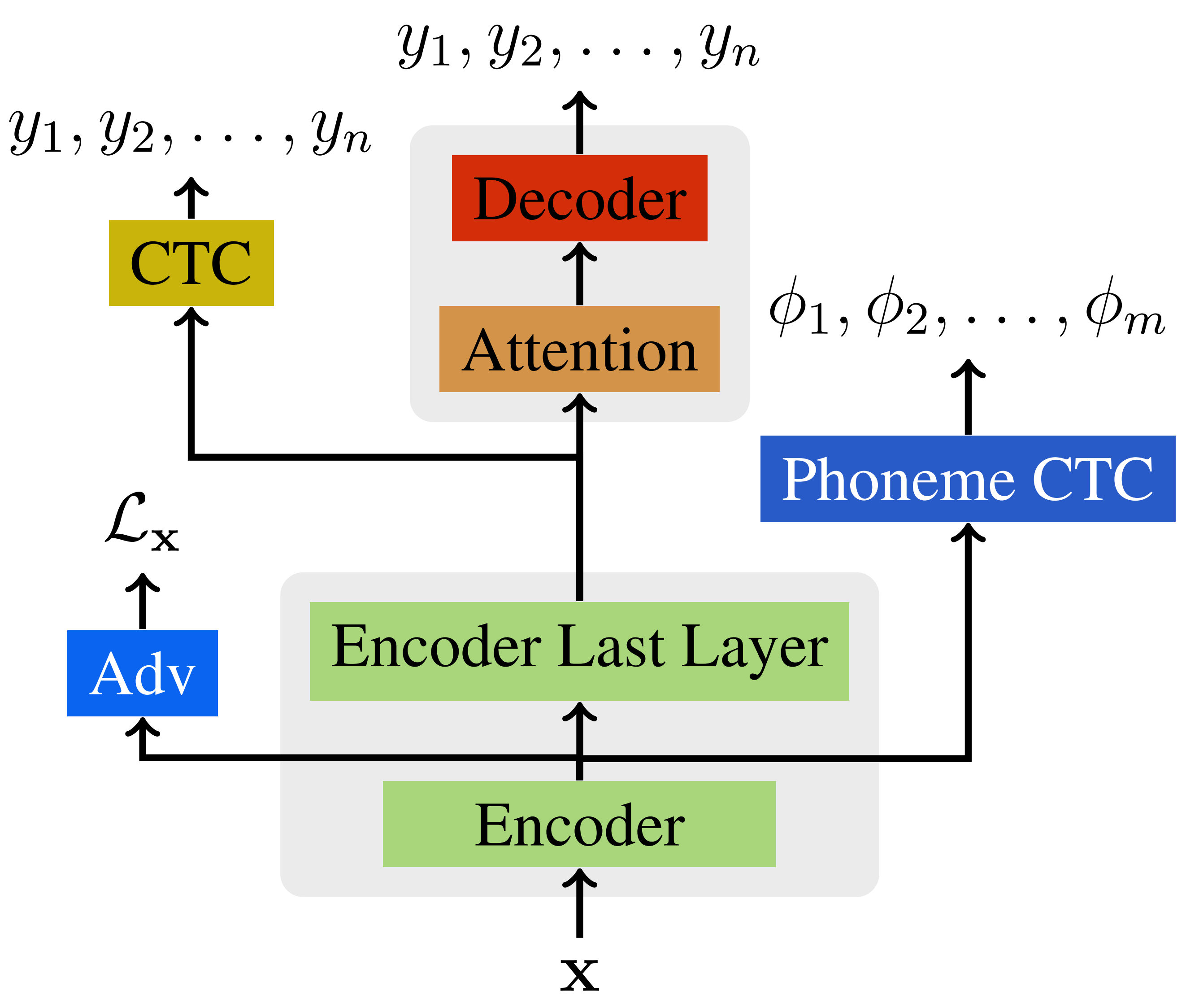
Massively multilingual adversarial ASR



- Pretrain multilingual ASR models using speech from as many as 100 languages!
- To encourage learning language-independent representations:
  - Context-independent phoneme sequence prediction
  - Domain-adversarial language classification objective to encourage language invariance

Image from: Adams et al.,Massively multingual adversarial ASR, 2019

Massively multilingual adversarial ASR



Comparion of pretrained models + auxiliary objectives

	MONO	QUE+CYR		PHONOLOGY		GEO		100-LANG	
		-	+phn+adv	-	+phn+adv	-	+phn+adv	-	+phn+adv
ayr	40.6	34.6	34.2 (-1.2%)	**33.9**	34.5 (+1.8%)	35.4	34.9 (-1.4%)	34.2	34.5 (+0.9%)
quh	14.8	14.9	**13.9** (-6.7%)	14.4	14.5 (+0.7%)	15.5	14.8 (-4.5%)	15.1	14.7 (-2.6%)
kek	23.9	24.8	23.7 (-4.4%)	24.8	24.5 (-1.2%)	23.0	**22.9** (-0.4%)	24.9	24.4 (-2.0%)
ixl	20.7	21.2	20.1 (-5.2%)	-	-	**19.7**	20.1 (+2.0%)	20.8	20.6 (-1.0%)
mlg	45.2	43.5	**41.4** (-4.8%)	43.2	41.7 (-3.5%)	43.3	42.2 (-2.5%)	44.4	42.2 (-5.0%)
ind	14.9	15.8	14.7 (-7.0%)	**13.7**	14.3 (+4.4%)	14.0	**13.7** (-2.1%)	14.7	14.2 (-3.4%)
kia	14.6	14.6	13.2 (-9.6%)	-	-	**12.1**	**12.1** (-0.0%)	14.4	13.0 (-9.7%)
swe	**20.5**	22.7	21.6 (-4.9%)	26.4	24.2 (-8.3%)	22.0	21.2 (-3.6%)	23.9	24.6 (+2.9%)
spn	14.5	19.7	14.4 (-26.9%)	13.9	13.8 (-0.7%)	13.1	**12.1** (-7.6%)	15.8	14.8 (-6.3%)
		Avg. rel. Δ:	(-7.8%)	Avg. rel. Δ:	(-1.0%)	Avg. rel. Δ:	(-2.3%)	Avg. rel. Δ:	(-2.9%)

Image from: Adams et al.,Massively multingual adversarial ASR, 2019