TRANSFER LEARNING OF LANGUAGE-INDEPENDENT END-TO-END ASR WITH LANGUAGE MODEL FUSION

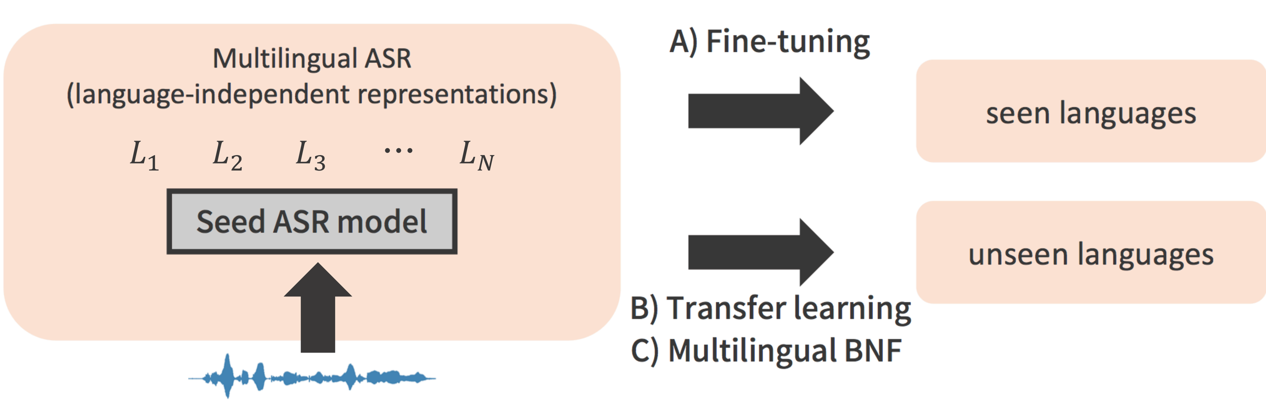
Hirofumi Inaguma[1] Jaejin Cho[2] Murali Karthick Baskar[3] Tatsuya Kawahara[1] Shinji Watanabe[2]
[1]Graduate School of Informatics, Kyoto University, Japan / [2]Johns Hopkins University, USA / [3]Brno University of Technology, Czech Republic

Summary

- Adapt language-independent sequence-to-sequence (S2S) ASR to low-resource languages (~50h)
- The language diversity is more important than the amount of training data
- The external RNNLM is integrated to the S2S model during adaptation (*LM fusion transfer*)
- Compared three *LM fusion transfer* methods
  1) Transfer learning + shallow fusion
  2) Deep fusion transfer
  3) Cold fusion transfer
- Cold fusion transfer is the most effective when the additional text is available
- Achieved the competitive performances to the BLSTM-HMM hybrid systems

Background: low-resource ASR

- Utilize data of other languages for data sparseness issue
  - A) Multi-task learning with other languages (multilingual training)
    → Further fine-tune to a particular language
  - B) Transfer learning from multilingual ASR (this work)
  - C) Adaptation with multilingual bottle-neck features (BNF)



- Goal: quick development of ASR systems for new languages
- Why End-to-End ASR?
  → Simplified training and decoding schemes (no need for lexicon per language)
- How to build the competitive systems to conventional hybrid systems?
  → Transfer learning from the well-trained language-independent ASR

Proposed method: LM fusion transfer

- Research question: Is linguistic context also helpful for adaptation to new languages?
  → Leverage the external monolingual RNNLM in target languages only in the adaptation stage

Adaptation scheme
1. Train character-level language-independent S2S ASR (unified vocabulary, 5353 classes)
2. Prepare the monolingual RNNLM on target languages
3. Copy all parameters from the language-independent S2S ASR
4. Integrate the external RNNLM during and/or after adaptation to target languages

d_u^{LM}: a hidden state of RNNLM
s_u^{S2S}: a hidden state of the decoder network

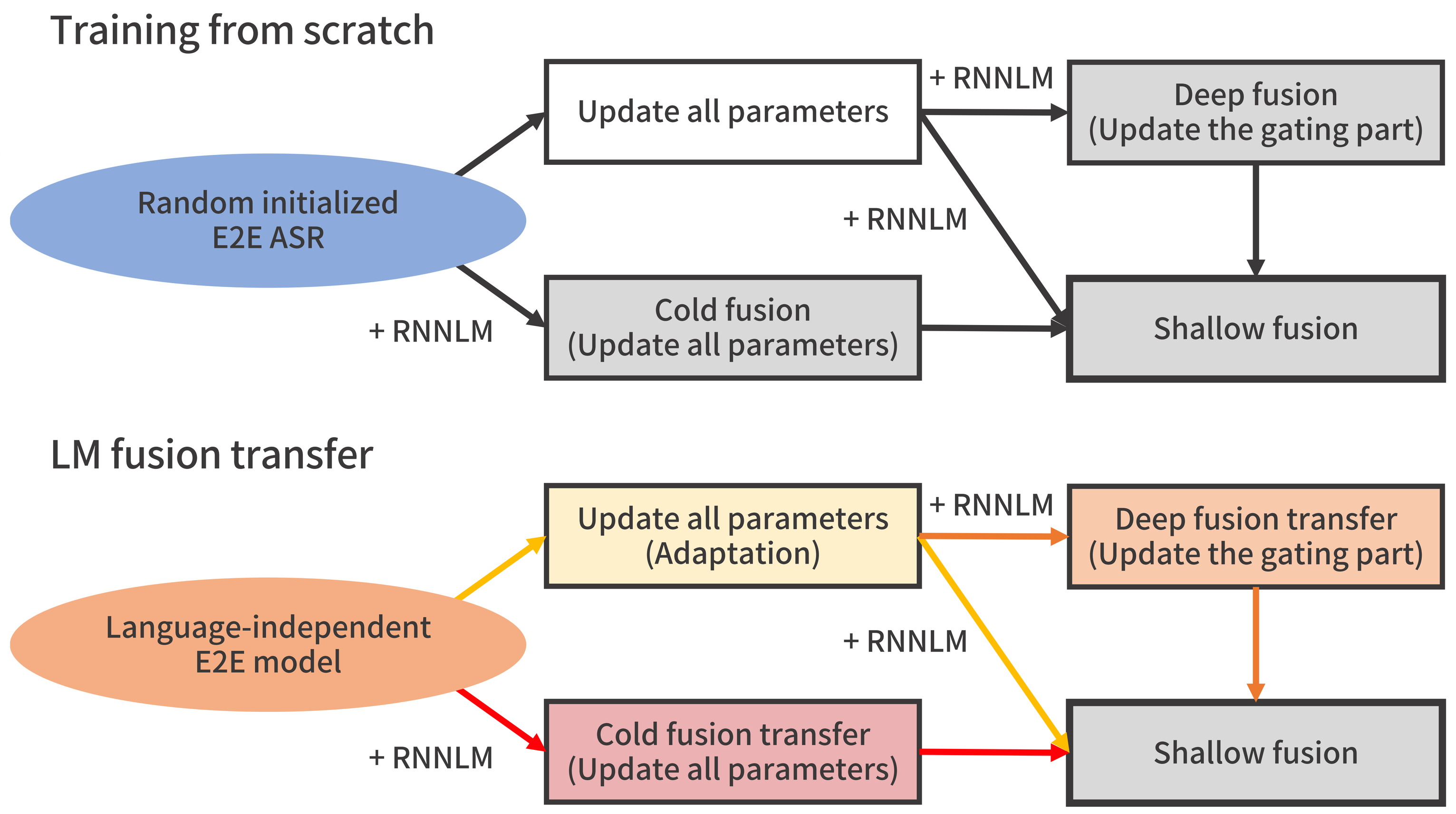
- LM fusion transfer
  - *Transfer + shallow fusion (SF)*
    - Interpolate RNNLM scores in the inference stage after adaptation

$$y^* = \underset{y\in\Omega^*}{\arg\max}\{\log P_{\mathrm{S2S}}(\boldsymbol{y}|\boldsymbol{x}) + \beta \log P_{\mathrm{LM}}(\boldsymbol{y})\}$$

  - *Cold fusion transfer (CF)*
    - The external RNNLM is integrated from the start point of adaptation

$$\boldsymbol{s}_u^{\mathrm{LM}} = \boldsymbol{W}^{\mathrm{LM}}\boldsymbol{d}_u^{\mathrm{LM}} + \boldsymbol{b}^{\mathrm{LM}}$$
$$\boldsymbol{g}_t = \sigma\left(\boldsymbol{W}^{\mathrm{g}}\left[\boldsymbol{s}_u^{\mathrm{S2S}}; \boldsymbol{s}_u^{\mathrm{LM}}\right] + \boldsymbol{b}^{\mathrm{g}}\right)$$
$$\boldsymbol{s}_u^{\mathrm{CF}} = \boldsymbol{W}^{\mathrm{CF}}\left[\boldsymbol{s}_u^{\mathrm{S2S}}; \boldsymbol{g}_t \odot \boldsymbol{s}_u^{\mathrm{LM}}\right] + \boldsymbol{b}^{\mathrm{CF}})$$
$$P_{\mathrm{S2S}}(\boldsymbol{y}|\boldsymbol{x}) = \mathrm{softmax}\left(\mathrm{ReLU}\left(\boldsymbol{W}^{\mathrm{out}}\boldsymbol{s}_u^{\mathrm{CF}} + \boldsymbol{b}^{\mathrm{o}}\right)\right)$$

  - *Deep fusion transfer (DF)*
    - The external RNNLM is integrated in the fine-tuning stage after adaptation



Experimental Evaluations

Data

multi15 (multi10+high2+target)

multi10 (BABEL)
Cantonese (126h), Bengali (55h), Pashto (70h), Turkish (68h), Vietnamese (78h), Haitian (60h), Tamil (62h), Kurmanji (37), Tokpisin (35h), Georgian (45h)

high2
Librispeech (English, 960h), CSJ (Japanese, 600h)

target languages (BABEL)
Assamese, Swahili, Lao, Tagalog, Zulu

Results

Result ①: Baseline monolingual systems for target 5 languages

Model	WER (%) Assamese (54h)	Swahili (39h)	Lao (58h)	Tagalog (75h)	Zulu (54h)
Old baseline [Cho 2018]	73.9	66.5	64.5	73.6	76.4
New baseline	64.5	56.6	56.2	56.4	69.5
+ large units (320→1024)	59.9	50.9	51.7	52.7	65.5
+ shallow fusion	**57.4**	**46.5**	**49.8**	**49.9**	**62.9**
BLSTM-HMM (monolingual)	49.1	38.3	45.7	46.3	61.1

+ VGG, 1L->2L decoder
Increasing the model capacity drastically improved the performance
Shallow fusion is always helpful though RNNLM is trained with small parallel data only
Competitive to BLSTM-HMM for Lao, Tagalog, and Zulu

Result ②: Comparison of seed language-independent models

Condition	Seed	hours	WER (%) Assamese (54h)	Swahili (39h)	Lao (58h)	Tagalog (75h)	Zulu (54h)
Unseen languages	multi10	643h	53.4	41.3	46.1	46.4	60.2
	high2	1,472h	57.8	45.0	48.6	49.4	61.9
	multi10+high2	2,115h	**53.2**	40.7	45.1	**45.3**	**58.5**
Seen languages	multi15	929h	53.4	**40.6**	**45.0**	46.1	58.8
	multi15 w/o fine-tune	929h	56.2	44.2	47.1	47.8	60.6

multi10 is almost sufficient for learning language-independent feature representation
The diversity of languages is more important than the total amount of training data

Result ③: LM fusion transfer (Full language pack (FLP): 50h speech data + FLP 50h text data)

Model		WER (%) Assamese (54h)	Swahili (39h)	Lao (58h)	Tagalog (75h)	Zulu (54h)
Transfer [Cho 2018]	SF	65.3	56.2	57.9	64.3	71.1
Scratch	-	59.9	50.9	51.7	52.7	65.5
	SF	**57.4**	46.5	**49.8**	**49.9**	62.9
	DF+SF	57.5	**46.4**	49.9	**49.9**	**62.6**
	CF+SF	57.5	47.3	50.0	50.2	62.9
Transfer (from *multi10*)	-	56.4	46.4	48.6	50.1	63.5
	SF	**53.4**	41.3	46.1	46.4	60.2
	DF+SF	53.5	**41.2**	46.2	**46.2**	59.9
	CF+SF	53.6	41.6	**45.9**	**46.2**	**59.5**

Proposed CF-transfer got some gains for 3 languages, but not significant because of using text in the parallel data only
Shallow fusion is more effective than when training from scratch
Outperformed the monolingual BLSTM-HMM system for Tagalog and Zulu, competitive for Lao

Result ④: LM fusion transfer (Limited language pack (LLP): 10h speech data + FLP 50h text data)

Model		LM data	WER (%) Assamese (8h)	Swahili (9h)	Lao (9h)	Tagalog (9h)	Zulu (9h)
Scratch	SF	-	Not converge				
Transfer (from *multi10*)	-	-	67.5	59.7	60.3	66.2	75.4
	SF	LLP (10h)	63.3	52.8	57.2	60.8	71.2
	DF+SF		68.0	52.4	57.3	60.7	70.9
	CF+SF		63.2	52.8	58.4	60.6	71.0
	SF	FLP (50h)	62.7	51.7	56.4	60.0	71.0
	DF+SF		66.8	50.7	56.1	60.0	69.9
	CF+SF		**61.7**	**50.3**	**56.0**	**57.9**	**69.8**

Linguistic context is helpful for adaptation when additional text data is available!

All LM fusion methods achieved a larger improvement even when RNNLM is trained with 10-hour data only
CF-transfer outperformed Transfer+SF on all 5 languages