On Language Model Integration for RNN Transducer based Speech Recognition

Wei Zhou, Zuoyun Zheng, Ralf Schlüter, Hermann Ney
zhou@cs.rwth-aachen.de
ICASSP 2022

Contents

Introduction

End-to-End (E2E) Speech Recognition

- great simplicity and state-of-the-art performance
- popular E2E approaches
  - connectionist temporal classification (**CTC**) [Graves & Fernández$^+$ 06]
  - recurrent neural network transducer (**RNN-T**) [Graves 12]
  - **attention**-based encoder-decoder models [Bahdanau & Chorowski$^+$ 16, Chan & Jaitly$^+$ 16]
- models only trained on paired audio-transcriptions

External Language Model (LM)

- much larger amount of text data
- possibly better-matched domain

$\rightarrow$ further boost the performance of E2E speech recognition

Proper LM Integration ?

Introduction

Previously,

- **shallow fusion (SF)** [Gulcehre & Firat+ 15a]: simple log-linear model combination
  - widely-used LM integration approach for E2E models
- other sophisticated approaches
  - e.g. deep fusion [Gülçehre & Firat+ 15b], cold fusion [Sriram & Jun+ 18]
  - higher complexity but not better than SF

Recently, **Internal Language Model (ILM)**

- RNN-T and attention models
  - context dependency directly included in the posterior distribution
  - implicitly learned sequence prior restricted to the audio transcription only
- **strong mismatch with the external LM**

$\rightarrow$ **limit the performance of LM integration such as simple SF**

3 Major Categories to handle ILM

- **ILM suppression**: suppress ILM in E2E model training
  - limiting context/model size [Zeineldeen & Glushko+ 21]
  - introducing an external LM at early stage [Michel & Schlüter+ 20]

- **ILM correction**: estimate and correct ILM from the posterior in decoding
  - various estimation methods [McDermott & Sak+ 19, Variani & Rybach+ 20, Meng & Parthasarathy+ 21, Zeyer & Merboldt+ 21, Zeineldeen & Glushko+ 21]
  - fits into a Bayesian interpretation

- **ILM adaptation**: adapt ILM on the same text data used by the external LM
  - train E2E models using text to speech [Deng & Zhao+ 21, Kurata & Saon+ 21, Rossenbach & Zeineldeen+ 21]
  - directly update partial model on text data [Pylkkönen & Ukkonen+ 21, Meng & Gaur+ 21]

Introduction

- ILM suppression
  - complexity: model/training modification
  - performs similarly well as ILM correction [Zeineldeen & Glushko+ 21]

- ILM adaptation
  - even higher complexity
  - usually aim at restricted application: no external LM
  - with external LM: ILM correction still needed [Deng & Zhao+ 21]

- **ILM correction**
  - **the most simple and effective LM integration approach**
  - also a better mathematical justification

  → major focus of this work: RNN-T

RNN-Transducer and Internal Language Model

RNN-T Recap

- sequence posterior

$$P_{\text{RNNT}}(a_1^S|X) = \sum_{y_1^{U=T+S}:\mathcal{B}^-1(a_1^S)} P_{\text{RNNT}}(y_1^U|h_1^T)$$

$$= \sum_{y_1^U:\mathcal{B}^-1(a_1^S)} \prod_{u=1}^{U=T+S} P_{\text{RNNT}}(y_u|\mathcal{B}(y_1^{u-1}), h_1^T)$$

  - a_1^S: output (sub)word sequence with $a \in V$
  - X: input acoustic feature sequence
  - $h_1^T = f^{\text{enc}}(X)$: encoder output
  - y_1^U: blank ϵ-augmented alignment sequence
  - unique mapping $\mathcal{B}(y_1^U) = a_1^S$: remove all ϵ

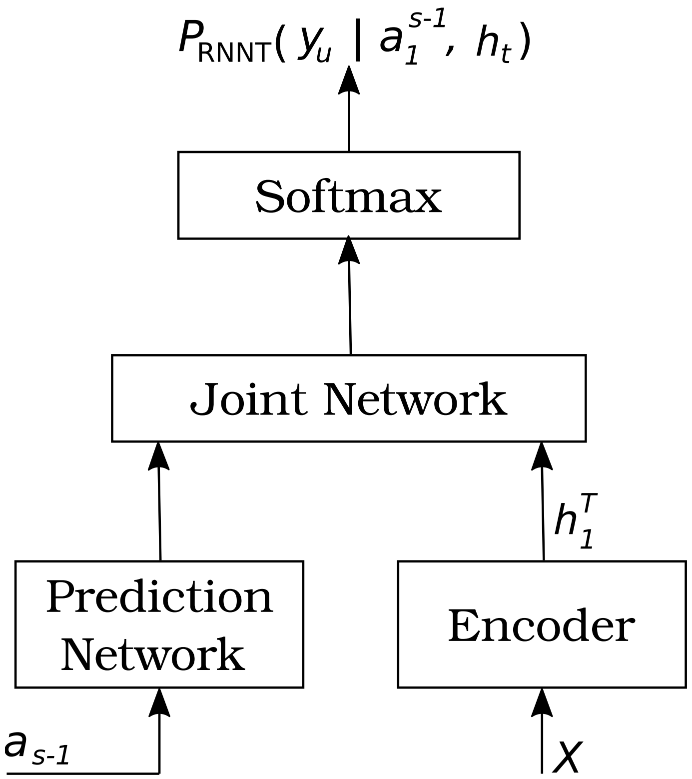
RNN-T Recap cont.

- neural network (NN): parameters θ_{RNNT}
  - encoder: f^{enc}
  - prediction network: f^{pred}
  - joint network: J

- lattice representation of RNN-T topology
  - y_1^{u-1}: a path reaching a node $(t, s-1)$

$$
\begin{aligned}
P_{\text{RNNT}}(y_u | \mathcal{B}(y_1^{u-1}), h_1^T) &= P_{\text{RNNT}}(y_u | a_1^{s-1}, h_t) \\
&= \text{Softmax}\left[J\left(f^{\text{pred}}(a_1^{s-1}), f_t^{\text{enc}}(X) \right) \right]
\end{aligned}
$$

  - y_1^u: reach $(t+1, s-1)$ if $y_u = \epsilon$, or (t, s) otherwise



RNN-Transducer and Internal Language Model

Maximum A Posteriori (MAP) Decoding

$$X \rightarrow \tilde{a}_1^{\tilde{S}} = \underset{a_1^S, S}{\arg\max} \; P(a_1^S|X)$$

- no external LM: simply plug in $P_{\text{RNNT}}(a_1^S|X)$
- **Bayesian framework: joint integration of the RNN-T model and an external LM**
  - Bayes' theorem: modularized components

$$X \rightarrow \tilde{a}_1^{\tilde{S}} = \underset{a_1^S, S}{\arg\max} \; P(a_1^S) \cdot P(X|a_1^S) = \underset{a_1^S, S}{\arg\max} \; P_{\text{LM}}^{\lambda_1}(a_1^S) \cdot \frac{P_{\text{RNNT}}(a_1^S|X)}{P_{\text{RNNT-ILM}}^{\lambda_2}(a_1^S)}$$

  - $P_{\text{RNNT-ILM}}$: ILM (sequence prior) implicitly learned and contained in P_{RNNT} $\rightarrow$ **ILM correction**
  - λ_1 and λ_2: scales applied in practice
    - shallow fusion (SF): omit $P_{\text{RNNT-ILM}}$ with $\lambda_2 = 0$

RNN-Transducer and Internal Language Model

ILM Estimation

- exact $P_{\text{RNNT-ILM}}$: intractable marginalization

$$P_{\text{RNNT-ILM}}(a_1^S) = \sum_X P_{\text{RNNT}}(a_1^S|X)P(X)$$

  $\rightarrow$ **approximation: estimated P_{ILM}**

- 2 major trends of estimation
  - statistics of the acoustic training transcription
    $\rightarrow$ **density ratio** [McDermott & Sak+ 19]: train a separate P_{ILM} on audio transcription
  - more consistent with the P_{RNNT} computation
    $\rightarrow$ **partially reuse the RNN-T NN for computing P_{ILM}**

RNN-Transducer and Internal Language Model

ILM Estimation: reuse RNN-T NN

$$P_{\text{ILM}}(a_1^S) = \prod_{s=1}^{S} P_{\text{ILM}}(a_s|a_1^{s-1}) = \prod_{s=1}^{S} P'(a_s|a_1^{s-1}, h')$$

$$P'(a_s|a_1^{s-1}, h') = \frac{P_{\text{RNNT}}(a_s|a_1^{s-1}, h')}{1 - P_{\text{RNNT}}(\epsilon|a_1^{s-1}, h')} = \text{Softmax}\left[J_{\backslash\epsilon}\left(f^{\text{pred}}(a_1^{s-1}), h'\right)\right]$$

- h': some global representation
- P': defined over V $\rightarrow$ same form as P_{RNNT}: usually over $V \cup \{\epsilon\}$
- simple renormalization
  - instead of separate ϵ distribution in P_{RNNT}: hybrid autoregressive transducer (HAT) [Variani & Rybach+ 20]
- $J_{\backslash\epsilon}$: joint network J excluding the ϵ logit output

RNN-Transducer and Internal Language Model

ILM Estimation: reuse RNN-T NN cont.

$$P_{\text{ILM}}(a_1^S) = \prod_{s=1}^{S} P'(a_s|a_1^{s-1}, h')$$

$$P'(a_s|a_1^{s-1}, h') = \text{Softmax}\left[J_{\setminus\epsilon}\left(f^{\text{pred}}(a_1^{s-1}), h'\right)\right]$$

1. $\mathbf{h'_{zero}}$: $h' = \vec{0}$ [Variani & Rybach$^+$ 20, Meng & Parthasarathy$^+$ 21]
2. $\mathbf{h'_{avg}}$: $h' = \text{mean}(h_1^T)$ [Zeyer & Merboldt$^+$ 21, Zeineldeen & Glushko$^+$ 21]
3. $\mathbf{h'_{a_1^{s-1}}}$: $h' = f_{\theta_{\text{ILM}}}(a_1^{s-1})$ where $f_{\theta_{\text{ILM}}}$ is an additional NN
   - $h'_{\text{mini-LSTM}}$ [Zeineldeen & Glushko$^+$ 21]: $f_{\theta_{\text{ILM}}} = \text{embedding}_{\text{RNNT}} \circ \text{LSTM}_{50} \circ \text{linear}$
   - training $f_{\theta_{\text{ILM}}}$ on audio transcription: $\mathcal{L}_{\text{ILM}} = -\log P_{\text{ILM}}(a_1^S)$

     $\rightarrow$ combines advantages: transcription statistics + reuse partial RNN-T NN

Note: these h′-based ILM estimation approaches are based on fixed θ_{RNNT}

RNN-Transducer and Internal Language Model

ILM Training (ILMT)

- h'-based ILM approaches: use partial RNN-T NN for P_{ILM}
  $\rightarrow$ **include ILM into RNN-T model training stage: ILMT**

- multi-task training of all parameters including θ_{RNNT}

$$\begin{aligned}
\mathcal{L}_{\text{RNNT}} &= -\log P_{\text{RNNT}}(a_1^S|X) \\
\mathcal{L}_{\text{ILM}} &= -\log P_{\text{ILM}}(a_1^S) \\
\mathcal{L}_{\text{ILMT}} &= \mathcal{L}_{\text{RNNT}} + \alpha \mathcal{L}_{\text{ILM}}
\end{aligned}$$

  - α: scaling factor
  - originally for the h'_{zero} approach [Variani & Rybach$^+$ 20, Meng & Kanda$^+$ 21]
  - also applicable for the h'_{avg} and $h'_{a_1^{s-1}}$ approaches

RNN-Transducer and Internal Language Model

Note: quality of $P_{\text{ILM}} \rightarrow$ how well it matches $P_{\text{RNNT-ILM}}$

Exact-ILM

- recall HAT [Variani & Rybach+ 20]: h'_{zero}-based P_{ILM}

$$\begin{aligned} \text{if} \quad & J_{\backslash\epsilon}\left(f^{\text{pred}}(a_1^{s-1}), f_t^{\text{enc}}(X)\right) = J_{\backslash\epsilon}\left(f^{\text{pred}}(a_1^{s-1})\right) + J_{\backslash\epsilon}\left(f_t^{\text{enc}}(X)\right) \\ \text{then} \quad & P_{\text{RNNT-ILM}}(a_s|a_1^{s-1}) \propto \exp\left[J_{\backslash\epsilon}\left(f^{\text{pred}}(a_1^{s-1}), h'_{\text{zero}}\right)\right] \end{aligned}$$

- **extension**

$$\begin{aligned} \text{if} \quad & J_{\backslash\epsilon}\left(f^{\text{pred}}(a_1^{s-1}), f_t^{\text{enc}}(X)\right) = J'(a_1^{s-1}) + J_{\backslash\epsilon}\left(f_t^{\text{enc}}(X)\right) \\ \text{then} \quad & P_{\text{RNNT-ILM}}(a_s|a_1^{s-1}) \propto \exp\left[J'(a_1^{s-1})\right] \end{aligned} \quad (1)$$

  - J': any function with output size $|V|$ + independent of X
- **exact-ILM training: train J' to fulfill the assumption $\rightarrow$ exact ILM estimation**
  - $\mathcal{L}_{J'}$: cross-entropy (CE) loss over Eq. (1)
  - simplification: Viterbi alignment of each X + only those h_t where a_s occurs

RNN-Transducer and Internal Language Model

Exact-ILM cont.

- $h'_{a_1^{s-1}}$**-based** P_{ILM} **+ exact-ILM training**

$$P_{\text{ILM}}(a_1^S) = \prod_{s=1}^{S} P'(a_s|a_1^{s-1}, h'_{a_1^{s-1}})$$

$$P'(a_s|a_1^{s-1}, h'_{a_1^{s-1}}) = \text{Softmax}\Big[\underbrace{J_{\backslash\epsilon}\big(f^{\text{pred}}(a_1^{s-1}), f_{\theta_{\text{ILM}}}(a_1^{s-1})\big)}_{J'(a_1^{s-1})}\Big]$$

  - train $f_{\theta_{\text{ILM}}}$ (fixed θ_{RNNT})

$$\mathcal{L}_{\text{ILM}}^{\text{exact}} = \mathcal{L}_{\text{ILM}} + \alpha \mathcal{L}_{J'}$$

  $\rightarrow$ **theoretical justification:** $P_{\textbf{RNNT-ILM}}(a_s|a_1^{s-1}) \propto \exp\left[J'(a_1^{s-1})\right]$

- other possibilities: e.g. joint training $\mathcal{L}_{\text{RNNT}} + \alpha \mathcal{L}_{J'}$ of all parameters
  - additionally force the model to better fulfill the assumption $\rightarrow$ exact ILM estimation

RNN-Transducer and Internal Language Model

Decoding Interpretation: why improvement with ILM correction ?

$$X \rightarrow \tilde{a}_1^{\tilde{S}} = \arg\max_{a_1^S, S} \; P_{\text{LM}}^{\lambda_1}(a_1^S) \cdot \frac{P_{\text{RNNT}}(a_1^S|X)}{P_{\text{ILM}}^{\lambda_2}(a_1^S)}$$

$$= \arg\max_{a_1^S, S} \sum_{y_1^U : \mathcal{B}^{-1}(a_1^S)} \prod_{u=1}^{U} P_{\text{RNNT}}(y_u|\mathcal{B}(y_1^{u-1}), h_1^T) \cdot Q(y_u|\mathcal{B}(y_1^{u-1})) \qquad \text{scoring in search}$$

$$\text{with } Q(y_u|\mathcal{B}(y_1^{u-1})) = \begin{cases} 1, & y_u = \epsilon \\ \frac{P_{\text{LM}}^{\lambda_1}(y_u|\mathcal{B}(y_1^{u-1}))}{P_{\text{ILM}}^{\lambda_2}(y_u|\mathcal{B}(y_1^{u-1}))}, & y_u \neq \epsilon \end{cases}$$

R1. prior removal rebalances label distribution of P_{RNNT}
$\rightarrow$ **rely more on external LM for context modeling (desired)**

R2. division by P_{ILM} boosts the label probability against (usually high) blank probability
$\rightarrow$ **increase importance of external LM (λ_1) without suffering huge deletion errors**

- limitation of SF ($\lambda_2 = 0$)
- no need of decoding heuristics: length-reward ...
- however tuning effort in practice: large $\lambda_2 \rightarrow$ insertion/substitution errors

Experiments

Setup

In-Domain: 960h Librispeech [Panayotov & Chen+ 15]

- 5k acoustic data-driven subword modeling (ADSM) units [Zhou & Zeineldeen+ 21]
- strictly monotonic RNN-T ($U = T$) [Tripathi & Lu+ 19]
  - 50-dimensional gammatone features [Schlüter & Bezrukov+ 07]
  - NN structure
    - f^{enc}: 6×640 bidirectional-LSTM
    - subsample 4: 2 max-pooling in f^{enc}
    - f^{pred}: $\text{embedding}_{256} \circ 2 \times 640$ LSTM
    - J: $\text{linear}_{1024}\text{-tanh} \circ \text{linear} \rightarrow \text{softmax}$
  - 45 epochs on Librispeech $\rightarrow$ base model for all experiments
- external LM: 32-layer Transformer

Cross-Domain: TED-LIUM Release 2 (TLv2) [Rousseau & Deléglise+ 14]

- external LM: 4×2048 long short-term memory (LSTM)

Experiments

Setup cont.

- density ratio LM: same structure as f^{pred} + Librispeech transcription
- $h'_{a_1^{s-1}}$ **ILM approach:** $h'_{\text{mini-LSTM}}$ with $f_{\theta_{\text{ILM}}} = \text{embedding}_{\text{RNNT}} \circ \text{LSTM}_{50} \circ \text{linear}_{1280}\text{-tanh}$
  1. train $f_{\theta_{\text{ILM}}}$ with $\mathcal{L}_{\text{ILM}}$: 0.5-1 epoch on Librispeech transcription only
  2. train $f_{\theta_{\text{ILM}}}$ with $\mathcal{L}_{\text{ILM}}^{\text{exact}} = \mathcal{L}_{\text{ILM}} + \alpha\mathcal{L}_{J'}$: 0.5-1 epoch on Librispeech audio & transcription
     - Viterbi alignment using the base RNN-T model
     - α: 1.0 for in-domain evaluation and 2.0 for cross-domain evaluation
- ILMT: $\mathcal{L}_{\text{ILMT}} = \mathcal{L}_{\text{RNNT}} + \alpha\mathcal{L}_{\text{ILM}}$ with $\alpha = 0.2$
  - applied for h'_{zero}, h'_{avg} and $h'_{\text{mini-LSTM}}$ ILM approaches
  - initialize with base RNN-T model + fine-tune upto 10 epochs on Librispeech
  - $\mathcal{L}_{\text{ILM}}$ only relevant for f^{pred} and J: freeze f^{enc}
- alignment-synchronous decoding [Saon & Tüske+ 20]
  - score-based pruning + beam limit 128
  - no heuristic approach: effect of each LM integration method
  - scales optimized on dev sets

LM Integration Evaluation

Model Train	Evaluation	Librispeech WER [%] dev clean	Librispeech WER [%] dev other	Librispeech WER [%] test clean	Librispeech WER [%] test other	TLv2 WER [%] dev	TLv2 WER [%] test
$\mathcal{L}_{\text{RNNT}}$	no LM	3.3	9.7	3.6	9.5	19.8	20.3
	SF	2.0	5.1	2.2	5.5	15.5	16.4
	density ratio	1.9	4.8	2.1	5.2	14.1	15.0
	h'_{zero}	1.8	4.4	2.0	4.8	13.6	14.4
	h'_{avg}	1.8	4.4	2.0	4.9	13.5	14.6
$+\ \mathcal{L}_{\text{ILM}}$	$h'_{\text{mini-LSTM}}$	1.8	4.3	1.9	4.7	13.4	14.4
$+\ \mathcal{L}_{\text{ILM}}^{\text{exact}}$		**1.8**	**4.2**	**1.9**	**4.6**	**13.2**	**14.0**
$\mathcal{L}_{\text{ILMT}}$	h'_{zero}	1.8	4.4	2.0	4.8	13.3	14.2
	h'_{avg}	1.9	4.5	2.1	4.9	13.5	14.4
	$h'_{\text{mini-LSTM}}$	1.8	4.4	2.0	4.8	13.2	14.1

- external LM: significant gain
- ILM correction: further large improvement over SF
- h'-based approaches: better than density ratio
  - $h'_{\text{mini-LSTM}}$ $(h'_{a_1^{s-1}})$: best
- **proposed $\mathcal{L}_{\text{ILM}}^{\text{exact}}$: further improve $h'_{\text{mini-LSTM}}$**
  - also better than $\mathcal{L}_{\text{ILMT}}$
- $\mathcal{L}_{\text{ILMT}}$: little effect on Librispeech
  - decreasing $\mathcal{L}_{\text{ILM}}$: no improvement on the overall performance (HAT [Variani & Rybach+ 20])

Experiments

Verification: 2 decoding-perspective reasons for improvement with ILM correction

R1. rebalance label distribution: rely more on external LM for context modeling

R2. boost label probability: increase importance of external LM (λ_1) without huge deletion errors

- **individual effect of R2 without effect of R1**: SF + length reward
- **individual effect of R1 without effect of R2**: h'_{zero} + renorm-ϵ
  - for each $y_u \neq \epsilon$
    1. renormalization:

$$P_{\text{norm}}(y_u) = \frac{P_{\text{RNNT}}(y_u|\mathcal{B}(y_1^{u-1}), h_1^T)/P_{\text{ILM}}^{\lambda_2}(y_u|\mathcal{B}(y_1^{u-1}))}{\sum_{a \in V} P_{\text{RNNT}}(a|\mathcal{B}(y_1^{u-1}), h_1^T)/P_{\text{ILM}}^{\lambda_2}(a|\mathcal{B}(y_1^{u-1}))}$$

    2. modify probability for search:

$$\left(1 - P_{\text{RNNT}}(\epsilon|\mathcal{B}(y_1^{u-1}), h_1^T)\right) \cdot P_{\text{norm}}(y_u) \cdot P_{\text{LM}}^{\lambda_1}(y_u|\mathcal{B}(y_1^{u-1}))$$

  - restrict label probability w.r.t. ϵ + maintain rebalanced label distribution to some extent

Verification: 2 decoding-perspective reasons for improvement with ILM correction

R1. rebalance label distribution: rely more on external LM for context modeling

R2. boost label probability: increase importance of external LM (λ_1) without huge deletion errors

Evaluation	λ_1	λ_2	Librispeech dev-other			
			WER	Sub	Del	Ins
SF	0.61	0	5.1	3.8	0.9	0.4
+ length reward	0.65		4.8	3.8	0.5	0.5
h'_{zero} + renorm-ϵ	0.61	0.35	4.9	3.6	0.9	0.4
+ length reward	0.65		4.6	3.7	0.5	0.4
h'_{zero}	0.85	0.4	**4.4**	**3.5**	**0.5**	**0.4**
+ length reward	0.95		4.5	3.6	0.5	0.4

- SF + length reward: reduced deletion errors
  $\rightarrow$ **verify R2 without R1**
- h'_{zero} + renorm-ϵ: reduced substitution errors, but still high deletion errors
  $\rightarrow$ **verify R1 without R2**
- h'_{zero} + renorm-ϵ + length reward: **complementary**
- h'_{zero}: enlarge R1 and R2 with larger scales
  - further improvement
  - **length reward not needed**

Conclusion

- RNN-T LM integration: mismatch between external LM and ILM $\rightarrow$ ILM correction
- detailed formulation: various ILM correction-based methods in a common RNN-T framework
- decoding interpretation: 2 major reasons for performance improvement with ILM correction
  - experimentally verified with detailed analysis
- exact-ILM training framework: extension upon HAT
  - theoretical justification for different ILM approaches
- systematic comparison: in-domain Librispeech and cross-domain TLv2
  - $h'_{\text{mini-LSTM}}$ $(h'_{a_1^{s-1}})$: best
  - $+$ exact-ILM training: further improvement

Thank you for your attention

References

[Bahdanau & Chorowski+ 16] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, Y. Bengio.
End-to-End Attention-based Large Vocabulary Speech Recognition.
In *Proc. ICASSP*, pp. 4945–4949, 2016.

[Chan & Jaitly+ 16] W. Chan, N. Jaitly, Q. Le, O. Vinyals.
Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition.
In *Proc. ICASSP*, pp. 4960–4964, 2016.

[Deng & Zhao+ 21] Y. Deng, R. Zhao, Z. Meng, X. Chen, B. Liu, J. Li, Y. Gong, L. He.
Improving RNN-T for Domain Scaling Using Semi-Supervised Training with Neural TTS.
In *Proc. Interspeech*, pp. 751–755, 2021.

[Graves & Fernández+ 06] A. Graves, S. Fernández, F. J. Gomez, J. Schmidhuber.
Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks.
In *Proc. Int. Conf. on Machine Learning (ICML)*, pp. 369–376, 2006.

References

[Graves 12] A. Graves.
Sequence Transduction with Recurrent Neural Networks, 2012.
https://arxiv.org/abs/1211.3711.

[Gulcehre & Firat+ 15a] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, Y. Bengio.
On Using Monolingual Corpora in Neural Machine Translation, 2015.
http://arxiv.org/abs/1503.03535.

[Gülçehre & Firat+ 15b] Ç. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, Y. Bengio.
On Using Monolingual Corpora in Neural Machine Translation, 2015.
http://arxiv.org/abs/1503.03535.

[Kurata & Saon+ 21] G. Kurata, G. Saon, B. Kingsbury, D. Haws, Z. Tüske.
Improving Customization of Neural Transducers by Mitigating Acoustic Mismatch of Synthesized Audio.

References

In *Proc. Interspeech*, pp. 2027–2031, 2021.

[McDermott & Sak+ 19] E. McDermott, H. Sak, E. Variani.
A Density Ratio Approach to Language Model Fusion in End-to-End Automatic Speech Recognition.
In *IEEE ASRU*, pp. 434–441, 2019.

[Meng & Gaur+ 21] Z. Meng, Y. Gaur, N. Kanda, J. Li, X. Chen, Y. Wu, Y. Gong.
Internal Language Model Adaptation with Text-Only Data for End-to-End Speech Recognition, 2021.
https://arxiv.org/abs/2110.05354.

[Meng & Kanda+ 21] Z. Meng, N. Kanda, Y. Gaur, S. Parthasarathy, E. Sun, L. Lu, X. Chen, J. Li, Y. Gong.
Internal Language Model Training for Domain-Adaptive End-To-End Speech Recognition.
In *Proc. ICASSP*, pp. 7338–7342, 2021.

[Meng & Parthasarathy+ 21] Z. Meng, S. Parthasarathy, E. Sun, Y. Gaur, N. Kanda, L. Lu, X. Chen, R. Zhao, J. Li, Y. Gong.

References

Internal Language Model Estimation for Domain-Adaptive End-to-End Speech Recognition.
In *IEEE SLT*, pp. 243–250, 2021.

[Michel & Schlüter+ 20] W. Michel, R. Schlüter, H. Ney.
Early Stage LM Integration Using Local and Global Log-Linear Combination.
In *Proc. Interspeech*, pp. 3605–3609, 2020.

[Panayotov & Chen+ 15] V. Panayotov, G. Chen, D. Povey, S. Khudanpur.
Librispeech: An ASR corpus based on public domain audio books.
In *Proc. ICASSP*, pp. 5206–5210, 2015.

[Pylkkönen & Ukkonen+ 21] J. Pylkkönen, A. Ukkonen, J. Kilpikoski, S. Tamminen, H. Heikinheimo.
Fast Text-Only Domain Adaptation of RNN-Transducer Prediction Network.
In *Proc. Interspeech*, pp. 1882–1886, 2021.

[Rossenbach & Zeineldeen+ 21] N. Rossenbach, M. Zeineldeen, B. Hilmes, R. Schlüter, H. Ney.

References

Comparing the Benefit of Synthetic Training Data for Various Automatic Speech Recognition Architectures.
In *IEEE ASRU*, Cartagena, Colombia, Dec. 2021.

[Rousseau & Deléglise+ 14] A. Rousseau, P. Deléglise, Y. Estève.
Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks.
In *Proc. LREC*, pp. 3935–3939, 2014.

[Saon & Tüske+ 20] G. Saon, Z. Tüske, K. Audhkhasi.
Alignment-Length Synchronous Decoding for RNN Transducer.
In *Proc. ICASSP*, pp. 7804–7808, 2020.

[Schlüter & Bezrukov+ 07] R. Schlüter, I. Bezrukov, H. Wagner, H. Ney.
Gammatone Features and Feature Combination for Large Vocabulary Speech Recognition.
In *Proc. ICASSP*, pp. 649–652, 2007.

[Sriram & Jun+ 18] A. Sriram, H. Jun, S. Satheesh, A. Coates.

References

Cold Fusion: Training Seq2Seq Models Together with Language Models.
In *Proc. Interspeech*, pp. 387–391, 2018.

[Tripathi & Lu+ 19] A. Tripathi, H. Lu, H. Sak, H. Soltau.
Monotonic Recurrent Neural Network Transducer and Decoding Strategies.
In *IEEE ASRU*, pp. 944–948, 2019.

[Variani & Rybach+ 20] E. Variani, D. Rybach, C. Allauzen, M. Riley.
Hybrid Autoregressive Transducer (HAT).
In *Proc. ICASSP*, pp. 6139–6143, 2020.

[Zeineldeen & Glushko+ 21] M. Zeineldeen, A. Glushko, W. Michel, A. Zeyer, R. Schlüter, H. Ney.
Investigating Methods to Improve Language Model Integration for Attention-based Encoder-Decoder ASR Models.
In *Proc. Interspeech*, pp. 2856–2860, 2021.

[Zeyer & Merboldt+ 21] A. Zeyer, A. Merboldt, W. Michel, R. Schlüter, H. Ney.

References

Librispeech Transducer Model with Internal Language Model Prior Correction.
In *Proc. Interspeech*, 2021.

[Zhou & Zeineldeen+ 21] W. Zhou, M. Zeineldeen, Z. Zheng, R. Schlüter, H. Ney.
Acoustic Data-Driven Subword Modeling for End-to-End Speech Recognition.
In *Proc. Interspeech*, pp. 2886–2890, 2021.