Microsoft

COMBINATION OF END-TO-END AND HYBRID MODELS FOR SPEECH RECOGNITION

Jeremy Wong, Yashesh Gaur, Rui Zhao, Liang Lu, Eric Sun, Jinyu Li, and Yifan Gong

Microsoft Speech and Language Group

Microsoft

Data

- Training:
  - 75K hours from variety of Microsoft applications.

- Testing:
  - Average of 13 application scenarios (Cortana, far-field, ….).
  - Total 1.8M words, 260K utterances.

Model architectures

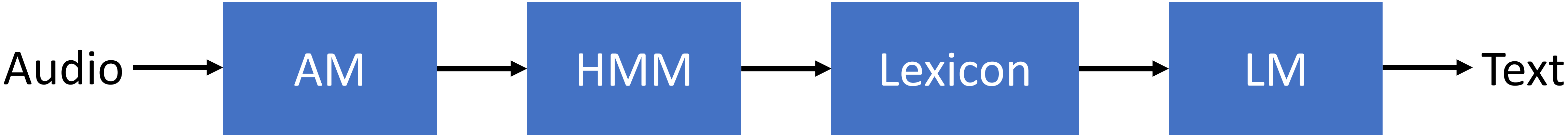
Hybrid model

$$P(\boldsymbol{\omega}_{1:L}|\mathbf{O}_{1:T}) \propto P^{\gamma}(\boldsymbol{\omega}_{1:L}) \sum_{\mathbf{s}_{1:T}\in\boldsymbol{\omega}_{1:L}} \prod_{t=1}^{T} \frac{P(s_t|\mathbf{o}_t)}{P(s_t)} P(s_t|s_{t-1})$$

- Language model

$$P(\boldsymbol{\omega}_{1:L}) = \prod_{l=1}^{L} P(\omega_l|\boldsymbol{\omega}_{l-n+1:l-1})$$

- Makes conditional independence assumptions.
- Uses external lexicon and language model.

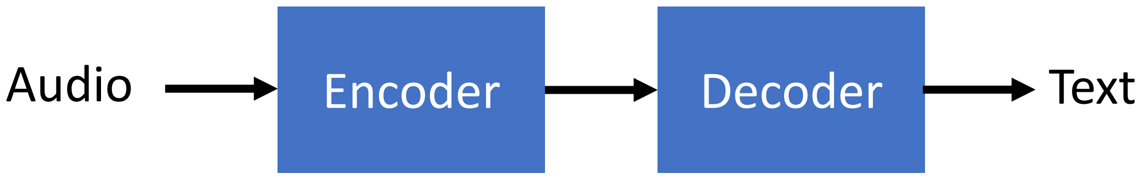


Model architectures

LAS model

$$P(\boldsymbol{\tau}_{1:J}|\mathbf{O}_{1:T}) = \prod_{j=1}^{J} P(\tau_j|\boldsymbol{\tau}_{1:j-1}, \mathbf{O}_{1:T})$$

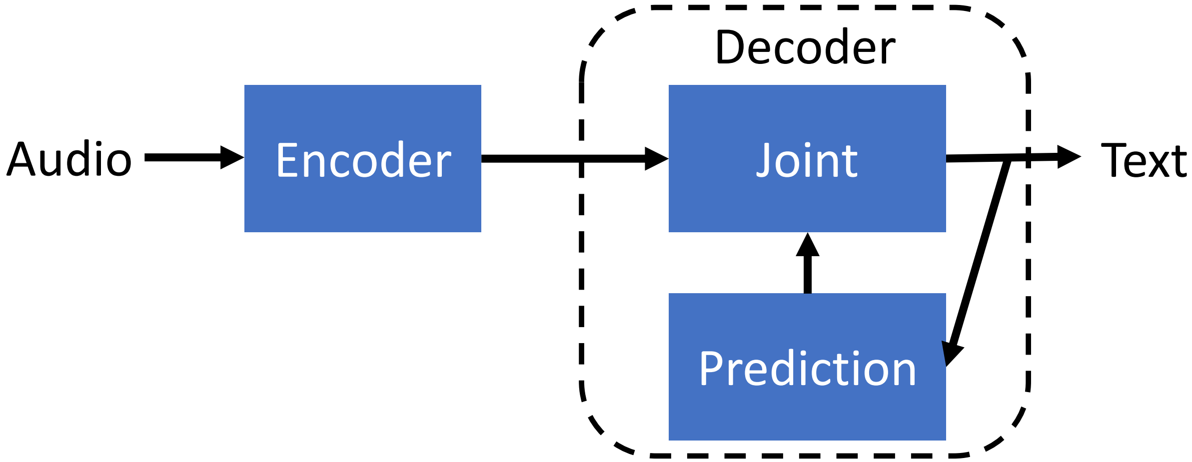
- No conditional independence assumption.
- All components jointly trained.
- Not frame-synchronous.



RNN-T model

$$P(\boldsymbol{\tau}_{1:J}|\mathbf{O}_{1:T}) = \sum_{\mathbf{s}_{1:T+J}\in\mathcal{B}(\boldsymbol{\tau}_{1:J},T)} \prod_{k=1}^{T+J} P(s_k|\mathbf{s}_{1:k-1}, \mathbf{O}_{1:T})$$

- No conditional independence assumption.
- All components jointly trained.
- Frame-synchronous.



Hypothesis-level model combination

- The models may behave differently and predict diverse error patterns.
- Combine the hypotheses together to correct each other's errors.
- Use MBR combination decoding.

$$\boldsymbol{\omega}^* = \underset{\boldsymbol{\omega}'}{\operatorname{argmin}} \sum_{m=1}^{M} \lambda_m \sum_{\boldsymbol{\omega} \in \mathbb{N}} \mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\omega}') \frac{P_m^{\kappa_m}(\boldsymbol{\omega}|\mathbf{O}_{1:T})}{\sum_{\breve{\boldsymbol{\omega}} \in \mathbb{N}} P_m^{\kappa_m}(\breve{\boldsymbol{\omega}}|\mathbf{O}_{1:T})}$$

- Only hypothesis posteriors are needed, not per-word scores.
- Performance depends on the accuracy of the hypothesis posteriors.

Bias toward short hypotheses

- LAS and RNN-T produce hypothesis posteriors that are biased toward short sequences.
- Alleviate using length normalisation.

$$\tilde{P}\left(\boldsymbol{\tau}_{1:J} \middle| \mathbf{O}_{1:T}\right) \propto P^{\frac{1}{J}}\left(\boldsymbol{\tau}_{1:J} \middle| \mathbf{O}_{1:T}\right)$$

Length norm	LAS WER (%)	Insertion (%)	Deletion (%)
no	10.40	0.79	4.82
yes	7.90	1.32	1.38

Model architectures

Hybrid model

$$P(\boldsymbol{\omega}_{1:L}|\mathbf{O}_{1:T}) \propto P^{\gamma}(\boldsymbol{\omega}_{1:L}) \sum_{\mathbf{s}_{1:T}\in\boldsymbol{\omega}_{1:L}} \prod_{t=1}^{T} \frac{P(s_t|\mathbf{o}_t)}{P(s_t)} P(s_t|s_{t-1})$$

- Language model

$$P(\boldsymbol{\omega}_{1:L}) = \prod_{l=1}^{L} P(\omega_l|\boldsymbol{\omega}_{l-n+1:l-1})$$

- Makes conditional independence assumptions.
- Uses external lexicon and language model.

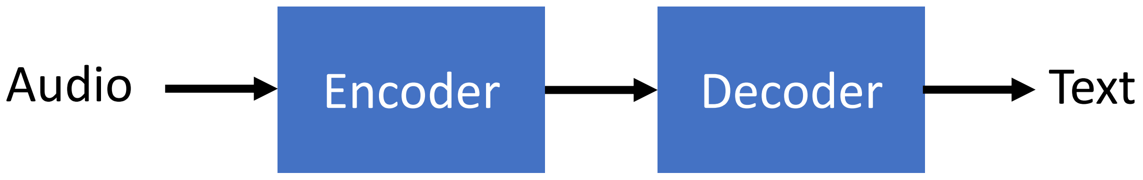
Audio → AM → HMM → Lexicon → LM → Text

Model architectures

LAS model

$$P(\boldsymbol{\tau}_{1:J}|\mathbf{O}_{1:T}) = \prod_{j=1}^{J} P(\tau_j|\boldsymbol{\tau}_{1:j-1}, \mathbf{O}_{1:T})$$

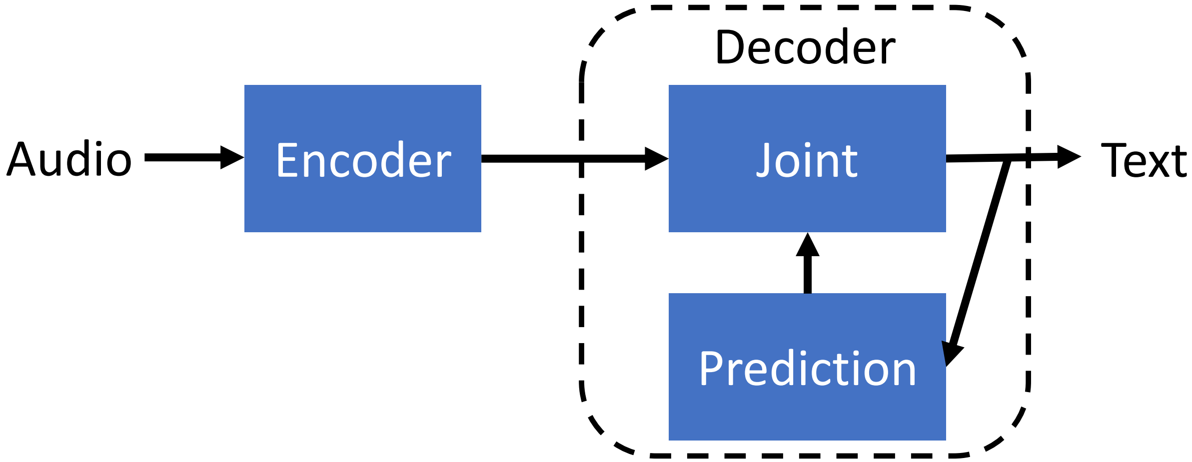
- No conditional independence assumption.
- All components jointly trained.
- Not frame-synchronous.



RNN-T model

$$P(\boldsymbol{\tau}_{1:J}|\mathbf{O}_{1:T}) = \sum_{\mathbf{s}_{1:T+J} \in \mathcal{B}(\boldsymbol{\tau}_{1:J}, T)} \prod_{k=1}^{T+J} P(s_k|\mathbf{s}_{1:k-1}, \mathbf{O}_{1:T})$$

- No conditional independence assumption.
- All components jointly trained.
- Frame-synchronous.



Microsoft

Bias toward short hypotheses

- LAS and RNN-T produce hypothesis posteriors that are biased toward short sequences.
- Alleviate using length normalisation.

$$\tilde{P}\left(\boldsymbol{\tau}_{1:J} \middle| \mathbf{O}_{1:T}\right) \propto P^{\frac{1}{J}}\left(\boldsymbol{\tau}_{1:J} \middle| \mathbf{O}_{1:T}\right)$$

Length norm	LAS WER (%)	Insertion (%)	Deletion (%)
no	10.40	0.79	4.82
yes	7.90	1.32	1.38

Microsoft

MBR training

- Can also alleviate bias by using discriminative training.
- Conditional maximum likelihood implicitly minimises alternative hypotheses through softmax.

$$\mathcal{F}_{\mathrm{CML}} = -\log P\left(\boldsymbol{\omega}^{\mathrm{ref}} \middle| \mathbf{O}_{1:T}\right)$$

- Minimum Bayes' risk explicitly minimises alternative hypotheses within criterion.

$$\mathcal{F}_{\mathrm{MBR}} = \sum_{\boldsymbol{\omega} \in \mathbb{N}} \mathcal{L}\left(\boldsymbol{\omega}, \boldsymbol{\omega}^{\mathrm{ref}}\right) \frac{P(\boldsymbol{\omega}|\mathbf{O}_{1:T})}{\sum_{\boldsymbol{\omega}' \in \mathbb{N}} P(\boldsymbol{\omega}'|\mathbf{O}_{1:T})}$$

- Length normalisation can be used inside MBR criterion.

$$\mathcal{F}_{\mathrm{MBR-LN}} = \sum_{\boldsymbol{\omega} \in \mathbb{N}} \mathcal{L}\left(\boldsymbol{\omega}, \boldsymbol{\omega}^{\mathrm{ref}}\right) \frac{P^{\frac{1}{|\boldsymbol{\omega}|}}(\boldsymbol{\omega}|\mathbf{O}_{1:T})}{\sum_{\boldsymbol{\omega}' \in \mathbb{N}} P^{\frac{1}{|\boldsymbol{\omega}'|}}(\boldsymbol{\omega}'|\mathbf{O}_{1:T})}$$

Microsoft

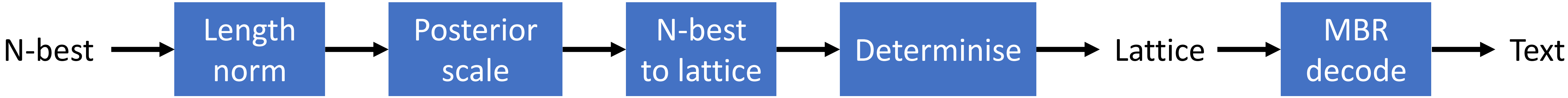
MBR training

Training	Decoding length norm	LAS WER (%)
$\mathcal{F}_{\text{CML}}$	no	10.40
	yes	7.90
$\mathcal{F}_{\text{MBR}}$	no	8.95
	yes	7.92
$\mathcal{F}_{\text{MBR-LN}}$	no	9.29
	yes	7.85

- MBR training reduces bias toward short hypotheses.

MBR decoding of end-to-end NN model

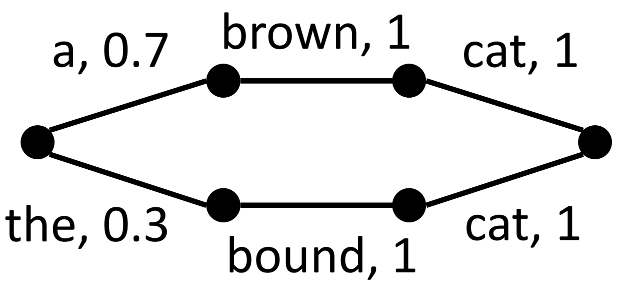
- Decoding process:



- Treat length-normalised scores as hypothesis posteriors.

- N-best to lattice conversion example:

a brown cat	0.7
the bound cat	0.3



MBR decoding of end-to-end NN model

Model	1-best WER (%)	MBR WER (%)
Hybrid	8.03	8.01
LAS	7.85	8.42
RNN-T	8.16	8.16

- N-best list size = 16.
- No significant gain from MBR decoding.

Microsoft

Model combination

- Hypothesis-level MBR combination.

Models	**WER (%)**	**Relative WERR (%)**
Hybrid	8.03	-
LAS	7.85	-
RNN-T	8.16	-
Hybrid + LAS	7.32	6.8
Hybrid + RNN-T	7.26	9.6
LAS + RNN-T	7.62	2.9
Hybrid + LAS + RNN-T	6.89	12.2

- Combination between different model architectures yields significant gains.

Microsoft

Model combination

- Compare combination methods for hybrid + LAS + RNN-T.

Combination method	**WER (%)**
1-best of merged N-best	7.59
ROVER	7.33
MBR	6.89

- MBR combination performs the best.

Conclusion

- Propose hypothesis-level combination between hybrid and end-to-end NN models.
- Length normalisation and MBR training can reduce bias toward short hypotheses.