ADVANCING ACOUSTIC-TO-WORD CTC MODEL

Jinyu Li, Guoli Ye, Amit Das, Rui Zhao, and Yifan Gong
Microsoft AI and Research, USA

Microsoft

1. Introduction

Acoustic-to-word models using CTC and whole word units are appealing, but suffer from having a closed vocabulary.

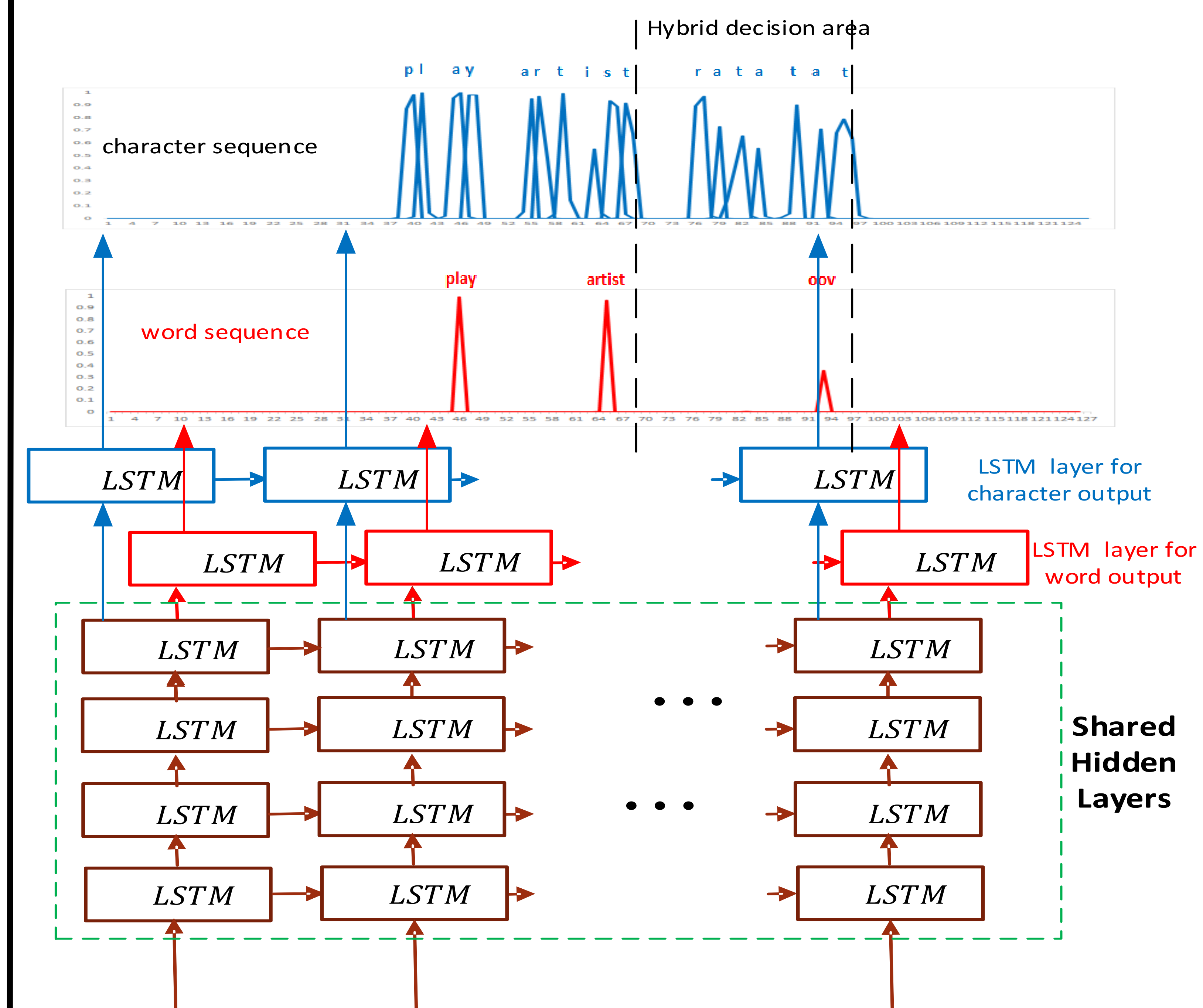
- Vocabulary is fixed at training time. All infrequent words are mapped into OOV, and cannot be modeled.
- Cannot easily handle emerging hot words.

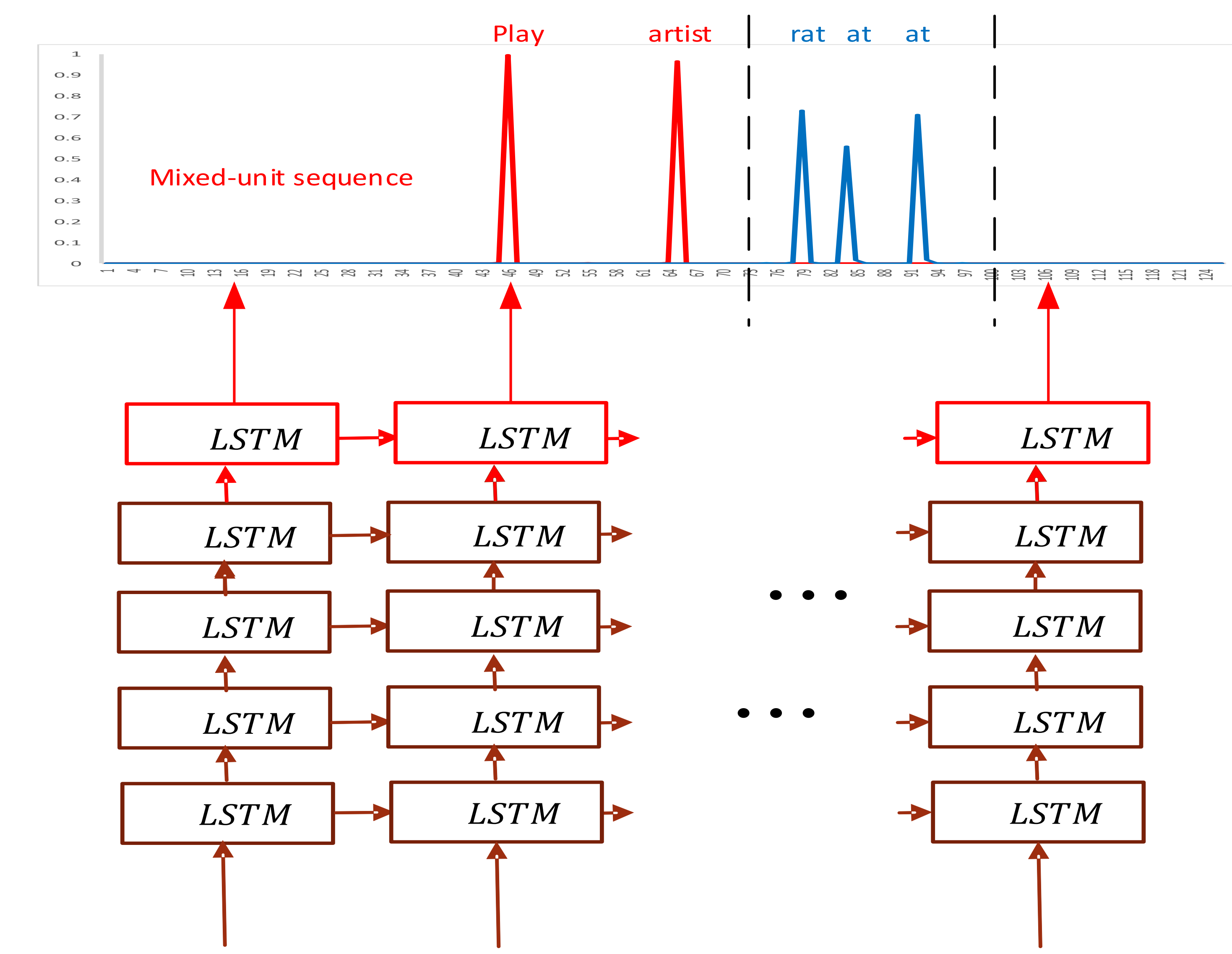
To address this OOV problem, we propose:

- Hybrid CTC: Simultaneously predicts both words and characters. Backs off to character outputs when the word model emits OOV tokens.
- CTC with mixed-unit: Decomposes all the OOV words into sequences of frequent words and letter n-gram units.

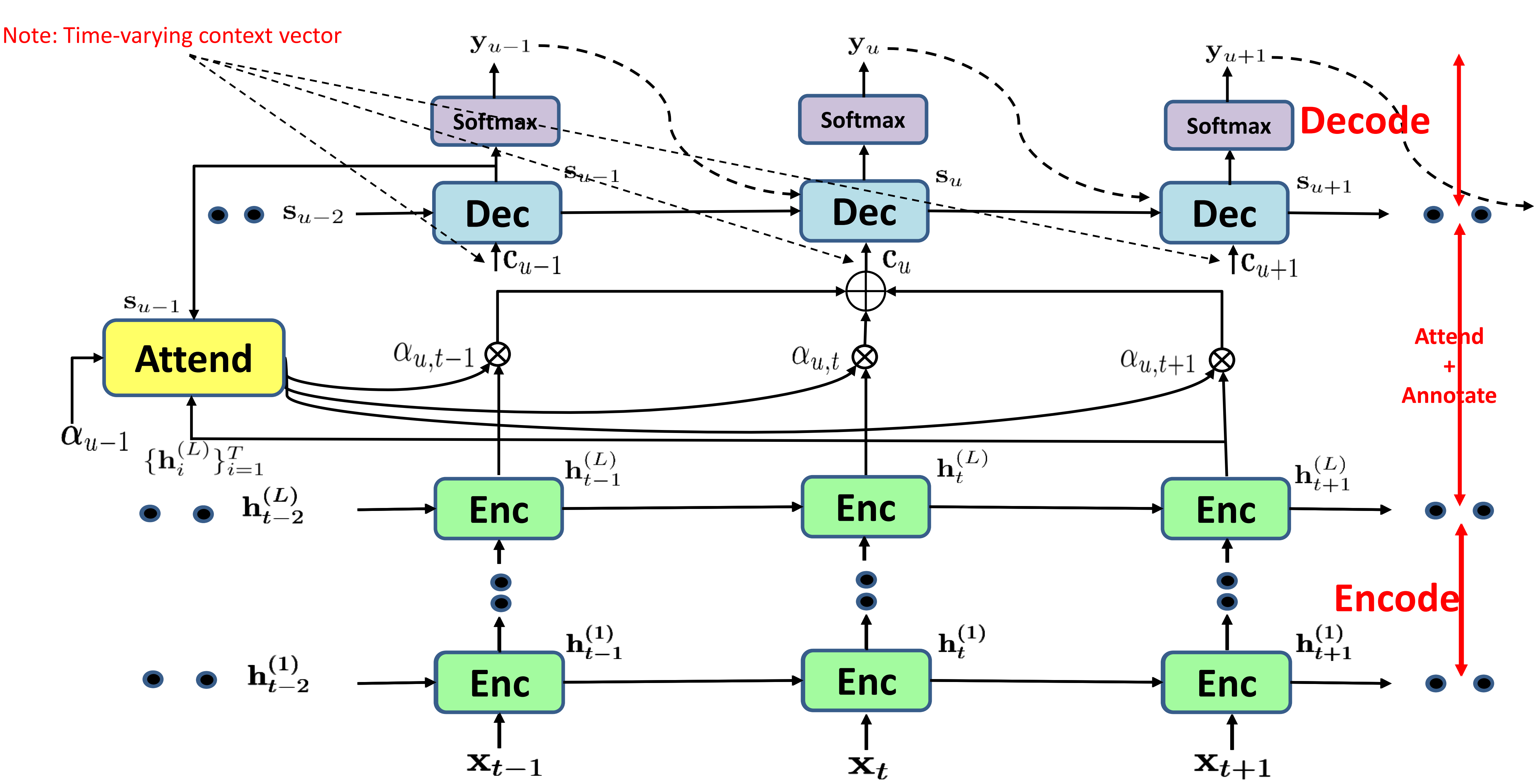
Combined with attention CTC, the final acoustic-to-word CTC ***beats*** the traditional CTC system with strong LM.

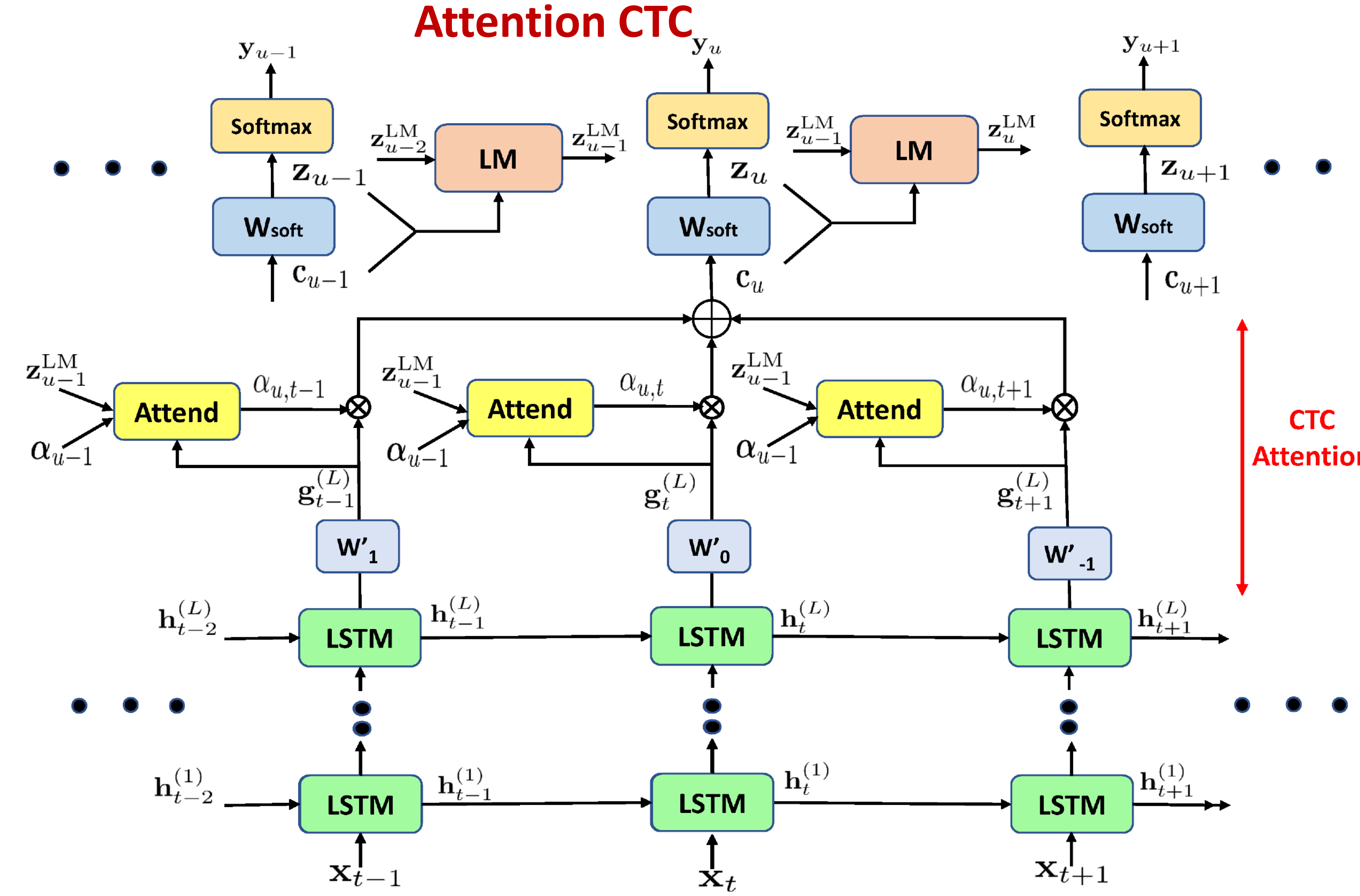
2. Hybrid CTC vs. Mixed-Unit CTC





3. Attention Encoder-Decoder vs. Attention CTC





ADVANCING ACOUSTIC-TO-WORD CTC MODEL

Jinyu Li, Guoli Ye, Amit Das, Rui Zhao, and Yifan Gong
Microsoft AI and Research, USA

Microsoft

Decomposition Type	Newyork	newyorkabc
All words: letter	n e w y o r k	n e w y o r k a b c
All words: letter 2-gram	ne wy or k	ne wy or ka bc
All words: letter 3-gram	new yor k	new yor kab c
All words: word	newyork	OOV
OOVs only: single-letter	newyork	n e w y o r k a b c
OOVs only: word+letter	newyork	newyork a b c
OOVs only: word+letter 3-gram	newyork	newyork abc

Table 1: Examples of how words are represented with different units

E2E Model	Vanilla	Attention	Attention 5-layer sharing	# of units
letter	17.54	14.30	16.74	*30*
letter 2-gram	15.37	12.16	14.00	*0.7k*
letter 3-gram	13.28	11.36	12.81	*8.9 k*

Table 2: WERs of letter-based CTC models

4. Experiments

- Training data:
  - 3400 hours of transcribed US-English Cortana audio
- Model:
  - 6-layer bi-directional LSTM, every layer has 512 memory units in each direction
  - Bi-directional CTC with CD-phone targets and 100M 5-gram: **9.28%** WER.
  - All end-to-end (E2E) models use **greedy decoding without LM.**
  - Bi-directional CTC with word targets gets 9.84% WER. OOV token contributes **1.87%** WER

E2E CTC Model	WER	# of units
Word-based	9.84	*27k*
Hybrid: Word-based + letter 2-gram Attention	9.66	*27k*
Hybrid: Word-based + letter 3-gram Attention	9.66	*35k*
Mixed (OOV: letter)	20.10	*27k*
Mixed (OOV: **word + letter**)	10.17	*27k*
Mixed (OOV: word + letter **2-gram**)	9.58	*27k*
Mixed (OOV: word + letter **3-gram**)	**9.32**	*33k*
Mixed (OOV: word + letter 3-gram) **Attention**	**8.65**	*33k*

Table 3: WERs of E2E CTCs

Attention CTC

$$e_{u,t} = \begin{cases} \mathbf{v}^T \tanh(\mathbf{U}\mathbf{z}_{u-1} + \mathbf{W}\mathbf{g}_t + \mathbf{b}), & \text{(content)} \\ \mathbf{v}^T \tanh(\mathbf{U}\mathbf{z}_{u-1} + \mathbf{W}\mathbf{g}_t + \mathbf{V}\mathbf{f}_{u,t} + \mathbf{b}), & \text{(hybrid)} \end{cases}$$

where, $\mathbf{f}_{u,t} = \mathbf{F} * \boldsymbol{\alpha}_{u-1}$

$$\alpha_{u,t} = \frac{\exp(e_{u,t})}{\sum_{t'=1}^{T} \exp(e_{u,t'})}, \quad t = [u-\tau, u+\tau]$$

Integration with implicit LM

$$\boldsymbol{\alpha}_u = \mathrm{Attend}(\mathbf{z}^{\mathrm{LM}}_{u-1}, \boldsymbol{\alpha}_{u-1}, \mathbf{g})$$

$$\mathbf{z}^{\mathrm{LM}}_{u-1} = \mathcal{H}(\mathbf{x}^{\mathrm{LM}}_{u-1}, \mathbf{z}^{\mathrm{LM}}_{u-2}), \quad \mathbf{x}^{\mathrm{LM}}_{u-1} = \begin{bmatrix} \mathbf{z}_{u-1} \\ \mathbf{c}_{u-1} \end{bmatrix}$$

Component-wise attention

$$\mathbf{e}_{u,t} = \begin{cases} \tanh(\mathbf{U}\mathbf{z}_{u-1} + \mathbf{W}\mathbf{g}_t + \mathbf{b}), & \text{(content)} \\ \tanh(\mathbf{U}\mathbf{z}_{u-1} + \mathbf{W}\mathbf{g}_t + \mathbf{V}\mathbf{f}_{u,t} + \mathbf{b}), & \text{(hybrid)} \end{cases}$$

where, $\mathbf{f}_{u,t} = \mathbf{F} * \boldsymbol{\alpha}_{u-1}$

$$\alpha_{u,t,j} = \frac{\exp(e_{u,t}(j))}{\sum_{t'=u-\tau}^{u+\tau} \exp(e_{u,t'}(j))}, \quad j = 1, \cdots, n.$$

- There is no explicit decoder in CTC network. Replace the decoder state $\mathbf{s}_{u-1}$ in Attention Encoder-Decoder with the logits $\mathbf{z}_{u-1}$ in Attention CTC.
- $\mathbf{z}^{\mathrm{LM}}_{u-1}$ captures long-term language information, but it is a pseudo-LM because of blanks in CTC.
- Instead of a single score per vector, we obtain a score for every component of the vector.

$$\mathbf{c}_u = \gamma \sum_{t=u-\tau}^{u+\tau} \boldsymbol{\alpha}_{u,t} \odot \mathbf{g}_t$$

5. Conclusions

- Advance acoustic-to-word CTC model with a **mixed-unit CTC**
  - Frequent word: model it with a unique output node.
  - OOV word: we decompose it into a sequence of frequent words and letter n-grams.
- Mixed-unit CTC is simpler and more effective than the 2-stage **hybrid CTC** which needs shared-hidden-layer to maintain the time synchronization of word outputs between the word-based and letter-based CTCs.
- The acoustic-to-word CTC with mixed-units reduces relative 5.28% WER from the vanilla word-based CTC, and reduces relative 12.09% WER if combined with the attention CTC.
- The final acoustic-to-word CTC outperforms the traditional context-dependent-phoneme CTC with strong LM and decoder by relative 6.79% WER reduction.
- It also provides more meaningful output without outputting any OOV token to distract users even if it cannot get the right words.
  - E.g., recognizes “text fabine” as “text fabian” and “call zubiate” as “call zubiat”, while the vanilla word-based CTC can only output “text OOV” and “call OOV”.