Unimodal Aggregation for CTC-based Speech Recognition

Ying Fang, Xiaofei Li*

Westlake University, Hangzhou, China

Introduction

Topic Non-autoregressive automatic speech recognition (NAR ASR)

AR methods vs. NAR methods

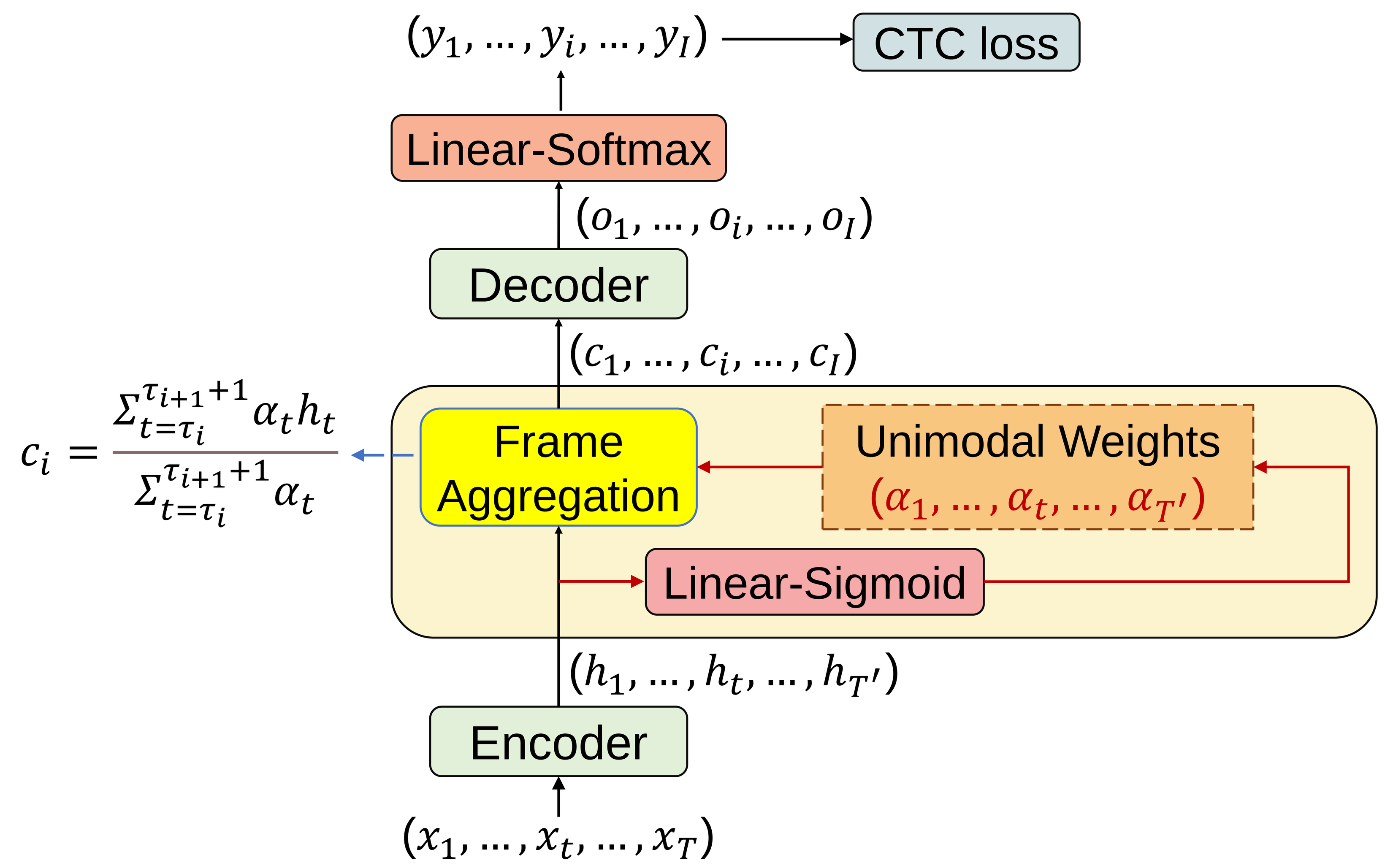
- AR: Attention mechanism ——Better performance, while serial and slow inference.
- NAR: CTC ——Reduced performance, but parallel and fast inference.

Proposed method Unimodal aggregation (UMA) , to segment and integrate the feature frames that belong to the same text token

Contributions
- Superior or comparable recognition performance to other advanced NAR methods on three Mandarin datasets.
- Shortens the sequence length, lower computational complexity.

Method

- **Encoder:** Transformer, Conformer, E-Branchformer, etc.
- **Unimodal aggregation module**
- **Decoder:** NAR self-attention network.



Denotation

- α_t: UMA weights, has first increasing and then decreasing pattern
- T', I: the sequence length before and after UMA
- τ_i: the time index of UMA valley, where $\alpha_t \leq \alpha_{t-1}$ and $\alpha_t \leq \alpha_{t-1}$

Results on HKUST

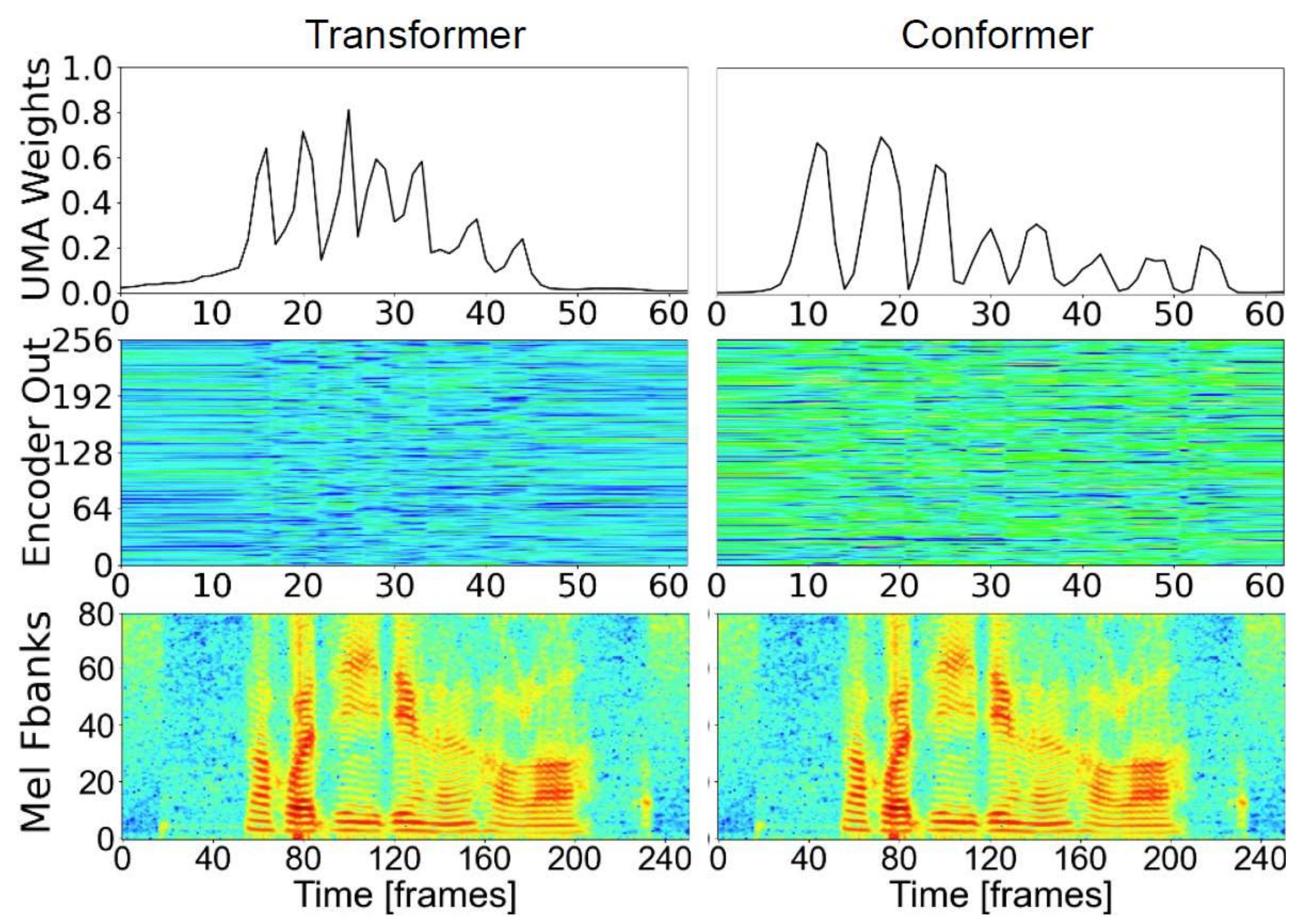
	Model	Transfomer				Conformer				E-Branchformer			
		sub	del	ins	CER	sub	del	ins	CER	sub	del	ins	CER
AR	Hybrid CTC/Attention	18.0	2.9	3.2	24.0	16.9	3.1	3.3	23.3	15.2	2.3	3.1	20.6
AR	+ beam search	15.9	2.8	2.8	**21.6**	15.7	2.5	3.0	**21.2**	14.1	2.3	2.8	**19.3**
NAR	CTC	18.4	3.0	3.3	24.7	17.3	2.8	3.2	23.2	16.0	2.6	2.9	21.6
NAR	Self-conditioned CTC	18.3	2.9	3.3	24.5	16.3	2.6	3.2	22.1	14.9	2.5	3.0	20.4
NAR	UMA (prop.)	15.9	6.5	2.6	25.0	15.6	2.7	3.2	21.4	14.1	3.4	2.6	20.1
NAR	+ self-condition	15.8	3.9	2.8	**22.6**	14.4	2.6	3.1	**20.0**	13.7	2.6	2.9	**19.2**

- May lead to extra deletion errors, adding self-conditioned layers can alleviate this
- Better encoder improve the quality of UMA weights

Conclusions

- UMA, a **simple yet effective** method for NAR ASR
- Learn better feature representation.
- Reduce the computation complexity
- Integrated with self-conditioned layers improves performance

Example



- Conformer encoder brings some time shifts, but its UMA weights are more discriminative.

Results on AISHELL-1/2

AISHELL-1 (178 hours)

	Model	dev	test	RTF	#Params(M)
AR	Hybrid (Conformer)	5.0	5.6	0.125	46.3
AR	+ beam search	**4.3**	**4.7**	0.461	46.3
NAR	LASO-large*	4.9	6.6	-	80.0
NAR	Paraformer*	4.6	5.2	-	-
NAR	CTC	5.6	6.1	0.052	50.4
NAR	Self-conditioned CTC	4.6	4.9	0.059	51.5
NAR	UMA (prop.)	4.5	4.8	0.039	42.6
NAR	+ self-condition	**4.4**	**4.7**	0.045	44.7

AISHELL-2 (1000 hours)

	Model	android	iOS	mic	RTF	#Params(M)
AR	Conformer	6.8	6.3	6.8	0.205	116.4
AR	+ beam search	**6.1**	**5.7**	**6.1**	0.954	116.4
NAR	LASO-large*	7.4	6.7	7.4	-	80.0
NAR	CIF+SAN*	6.2	5.8	6.3	-	-
NAR	UMA (prop.)	**6.0**	**5.3**	6.0	0.085	105.1
NAR	+ self-condition	**6.0**	**5.3**	**5.9**	0.098	110.4

- UMA outperforms all comparison NAR models.
- Achieves comparable performance with the hybrid CTC/attention+beam search
- Model size and RTF are both smaller than CTC

Email: fangying@westlake.edu.cn, lixiaofei@westlake.edu.cn Website: https://audio.westlake.edu.cn/ GitHub: https://github.com/Audio-WestlakeU/UMA-ASR