Tandem Long-Short Duration-based Modeling for Automatic Speech Recognition

Mengke Dalai, Meng Yan, Mihajilik Péter
BME TMIT SmartLabs

Workshop on SIGUL 2024

Speaker: Mengke Dalai

01 INTRODUCTION

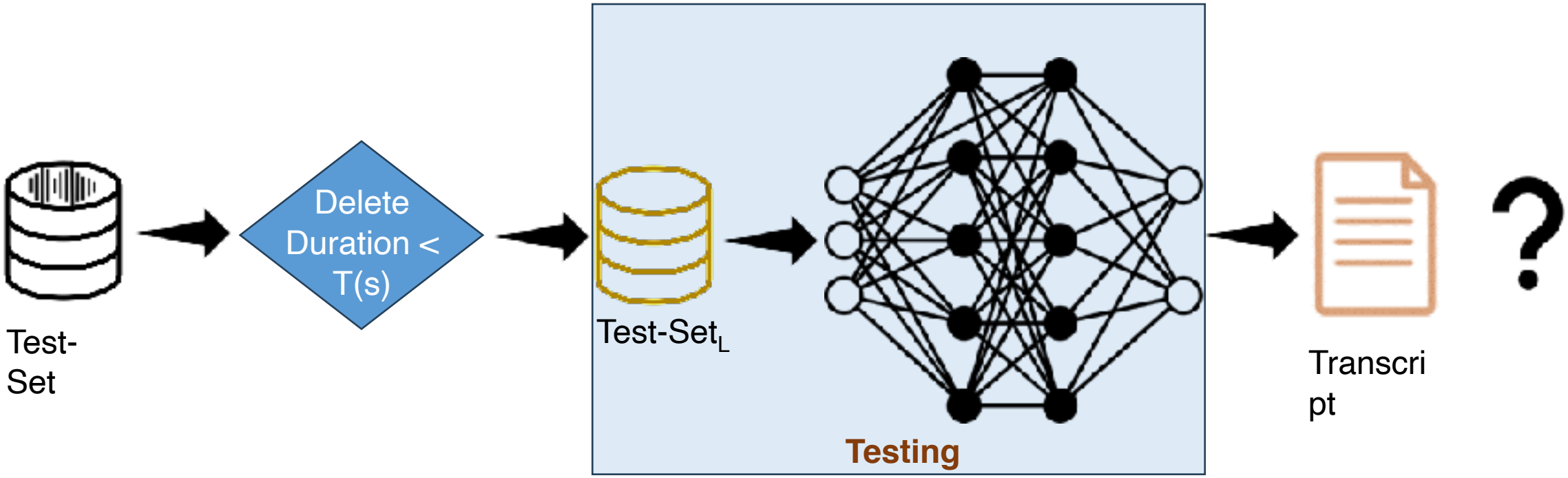
Challenges With Short Utterances in ASR Tasks

Lack of Context
The word "read" in "I read a book" can be past or present tense, the context helps identify the correct form. But like "Read it" do not provide such clue.

Homophones and Homonyms
A short utterance like "Right here," "right" could be misrecognized as "write" without additional context. The command "Go" might be confused with similar sounding words like "No".

Imbalance of Short and Long Utterances In Training Corpus
The lack of short utterances in the training corpus makes the model mainly learn the characteristics and patterns from long utterances during the training , lacks the learning for short utterances.

Preliminary Experiment: Delete Short Utterances From a Testing



Preliminary Experiment

Result After delete Short Utterances From Test-set

Duration	BEA-Base(eval-spont)%
T≥0s	25.42
T≥2s	24.85
T≥2.5s	24.70
T≥3.0s	24.72
T≥3.5s	**24.65**

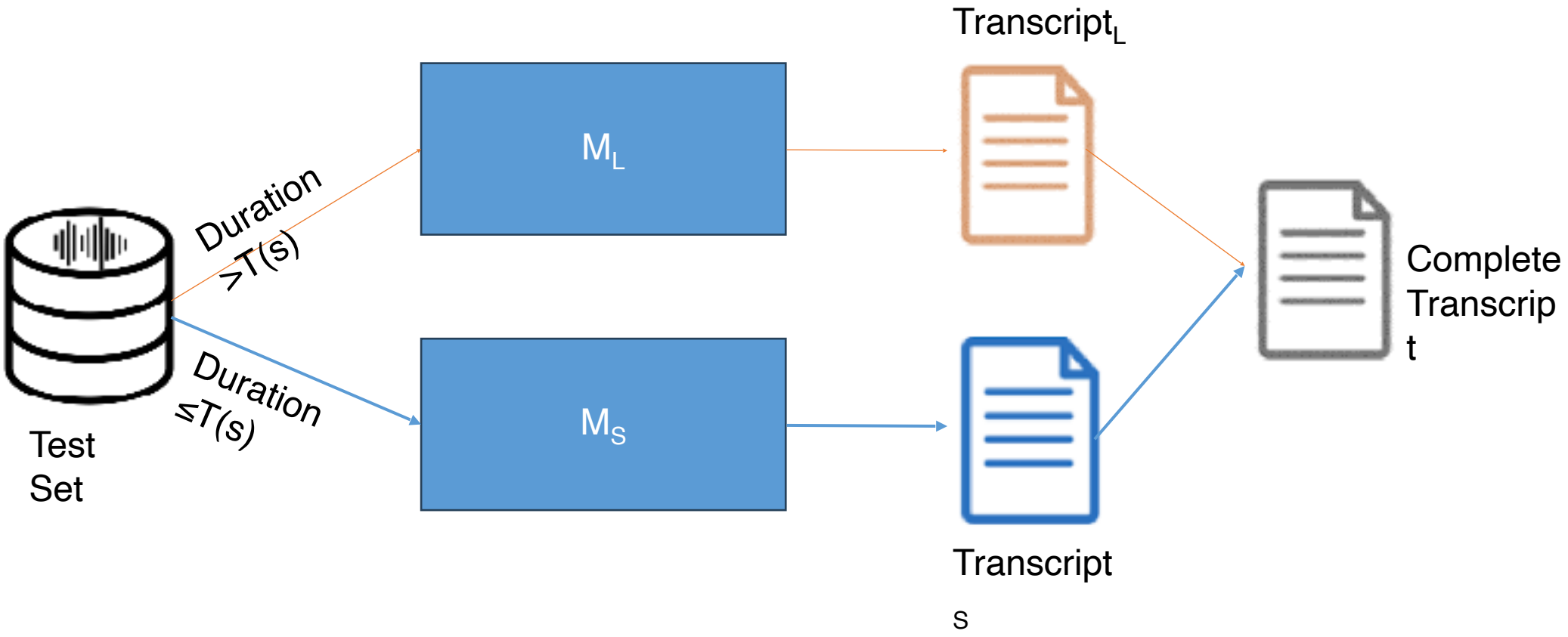
Table: Deletion result for BEA-Base

Duration	CV 15.0(test)%
T≥0s	23.72
T≥3.5s	23.47
T≥4.5s	23.25
T≥5.5s	23.05
T≥6.5s	**22.96**

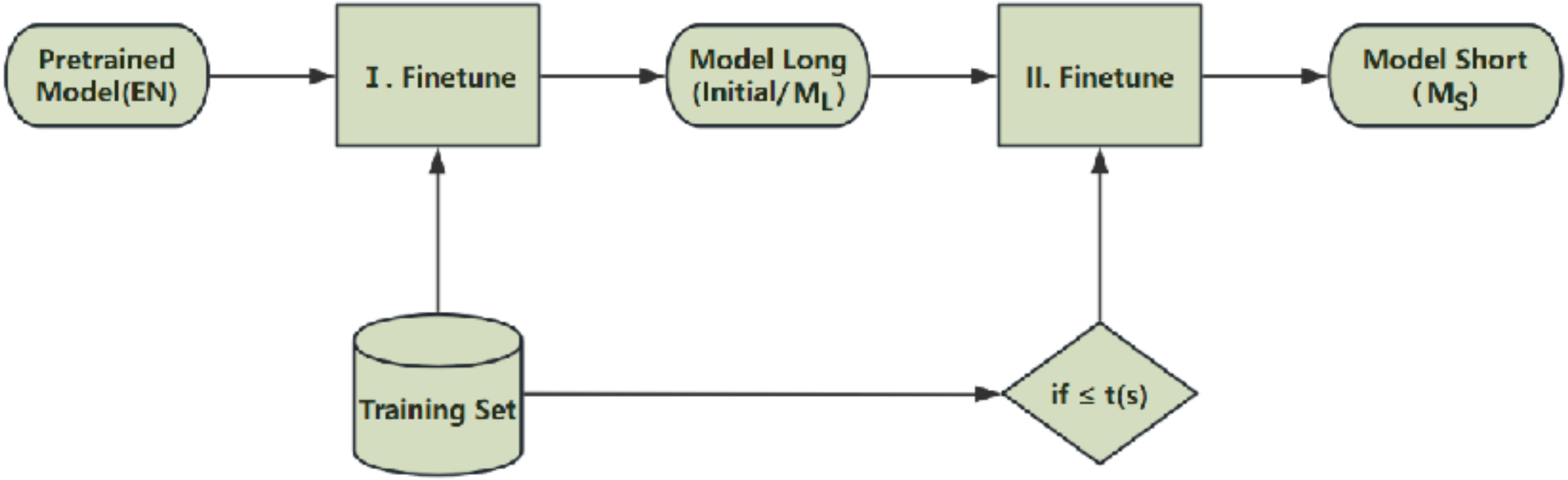
Table: Deletion result for CV

02 METHODOLOGY

How do we deal with short Utterances ?



How do we get M_S?



03 EXPERIMENTAL RESULTS

Short models improve the recognition of short sentences

	Split by Time(s)	ER_S on M_L(%)	ER_S on M_S(%)
BEA-Base	2.5	26.30	25.51 ↑
	3	25.67	25.10 ↑
	3.5	25.63	25.12 ↑
	Split by Time(s)	**ER_S on M_L(%)**	**ER_S on M_S(%)**
Common Voice(CV)	4.5	26.04	25.86 ↑
	5	25.46	24.46 ↑
	5.5	24.71	23.79 ↑
	6.0	24.34	23.17 ↑
	6.5	24.14	22.90 ↑

Comprehensive recognition improvement On BEA-Base

	T(s)	$N_{Error}W/N_T$ Word in Longer Utt Transcript	ER_L on M_L(%)	$N_{Error}W/N_T$ Word in Short Utt Transcript	ER_S on M_S(%)	1. Av. ER(%)	Baseline
BEA-Base	2.5	7083 / 28673	24.70	1660 / 6505	25.51	1. 24.85 ↑	25.42
	3.0	6291 / 25445	24.72	2443 / 9733	25.10	24.82 ↑	
	3.5	5484 / 22241	24.65	3249 / 12937	25.12	**24.82** ↑	

Comprehensive recognition improvement On CV

	T(s)	$N_{Error}W/N_T$ Word in Longer Utt Transcript	ER_L on M_L(%)	$N_{Error}W/N_T$ Word in Short Utt Transcript	ER_S on M_S(%)	Av. ER(%)	Baseline
Common Voice(CV)	4.5	16162 / 69513	23.25	3550 / 13726	25.86	23.68	23.72
	5.0	13786 / 59888	23.02	5712 / 23351	24.46	23.42	
	5.5	11436 / 49612	23.05	8001 / 33627	23.79	23.35	
	6.0	8895 / 38658	23.00	10330 / 44581	23.17	23.09	
	6.5	6775 / 29497	22.96	12310 / 53742	22.90	**22.92**	

CONCLUSION

What was the problem?
It was found that the automatic recognition for shorter utterances are generally more difficult than longer utterances.

How did we solve it?
We proposed a two-model corporation strategy, the longer utterances recognized by initial model(M_L), the shorter utterances recognized by the model(M_S), which is a further fine-tuning for initial model by short utterances.

What was the result?
These two models work together achieving a noticeable improvement in terms of WER on two public Hungarian datasets (BEA-Base, CV15.0).

Tandem Long-Short Duration-based Modeling for Automatic Speech Recognition

Thank You