AALBERT

Audio ALBERT: A Lite BERT for Self-supervised Learning of Audio Representation

Recap: ALBERT

- Factorize Embedding Matrix
- Share Parameters across layer
- Model Configuration between BERT

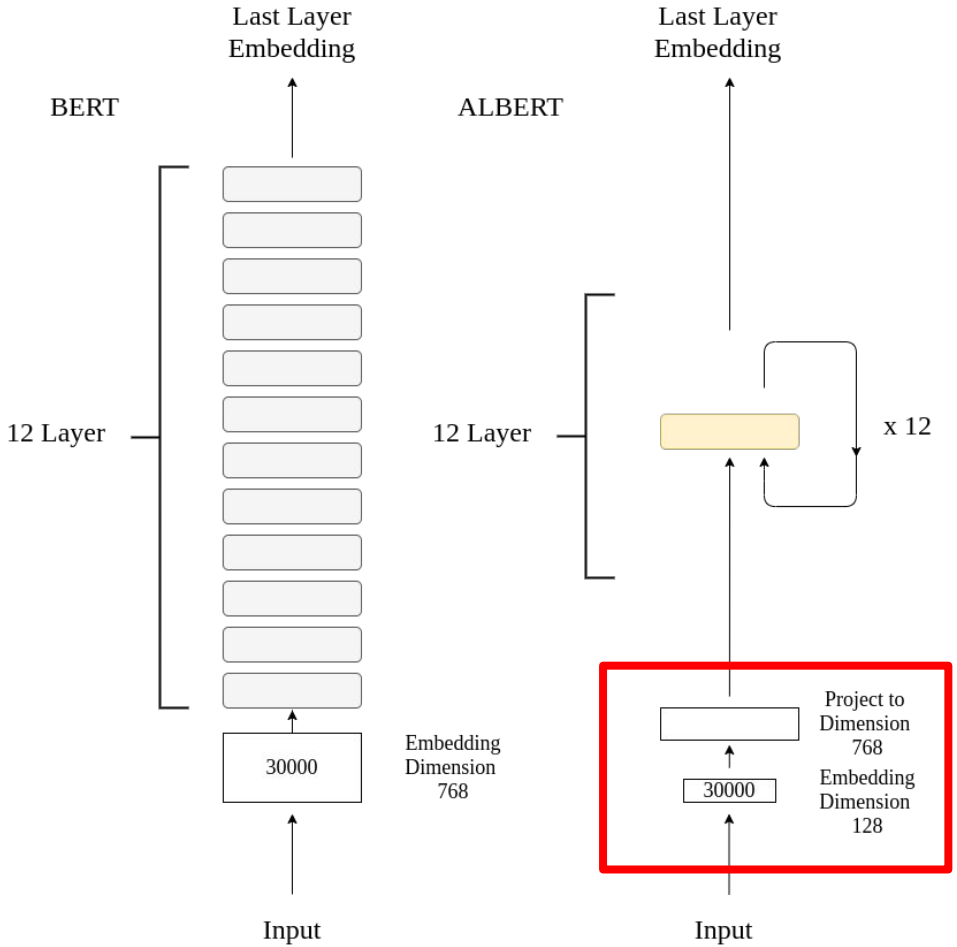
ALBERT

1.Factorize Embedding Matrix

Original BERT:
30000 x 768 = 23.04M

ALBERT:
30000 x128 = 3.8M
128 x 768 = 0.098M
Total: 3.898M

Reduce Parameters !

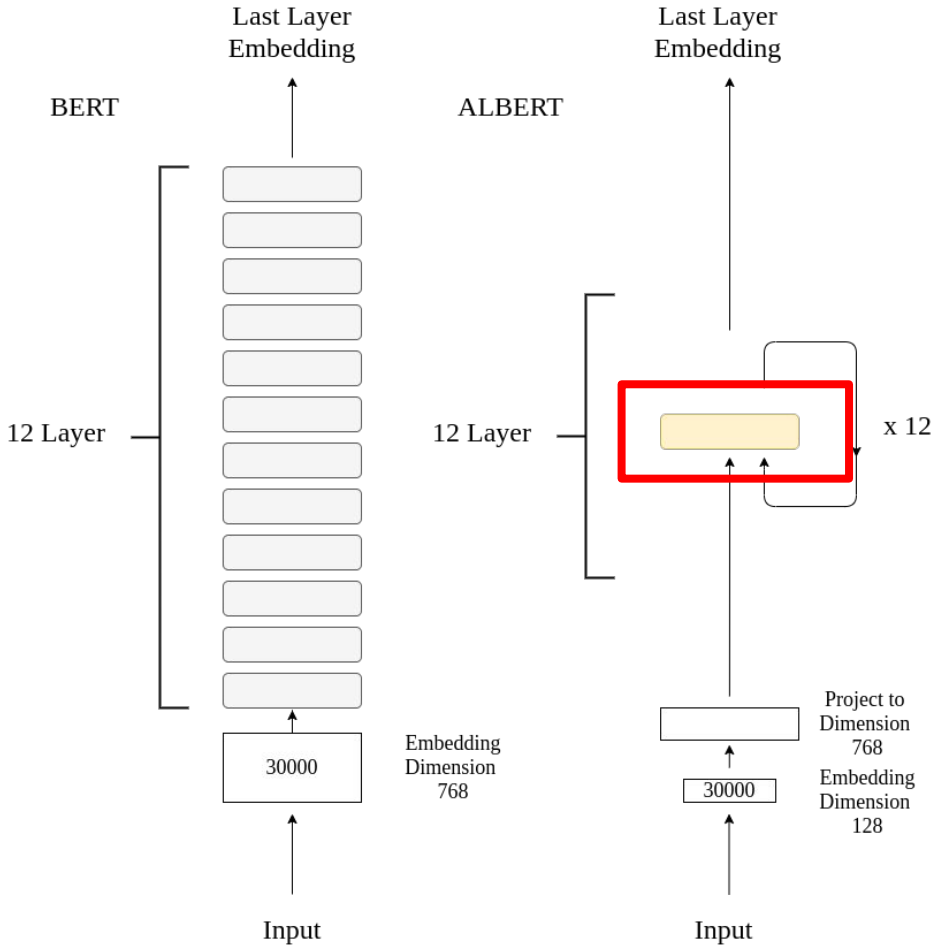


ALBERT

2.Shared Same Parameters across Layer

1/ 12 BERT Parameters on Layer

Reduce Parameters !!!



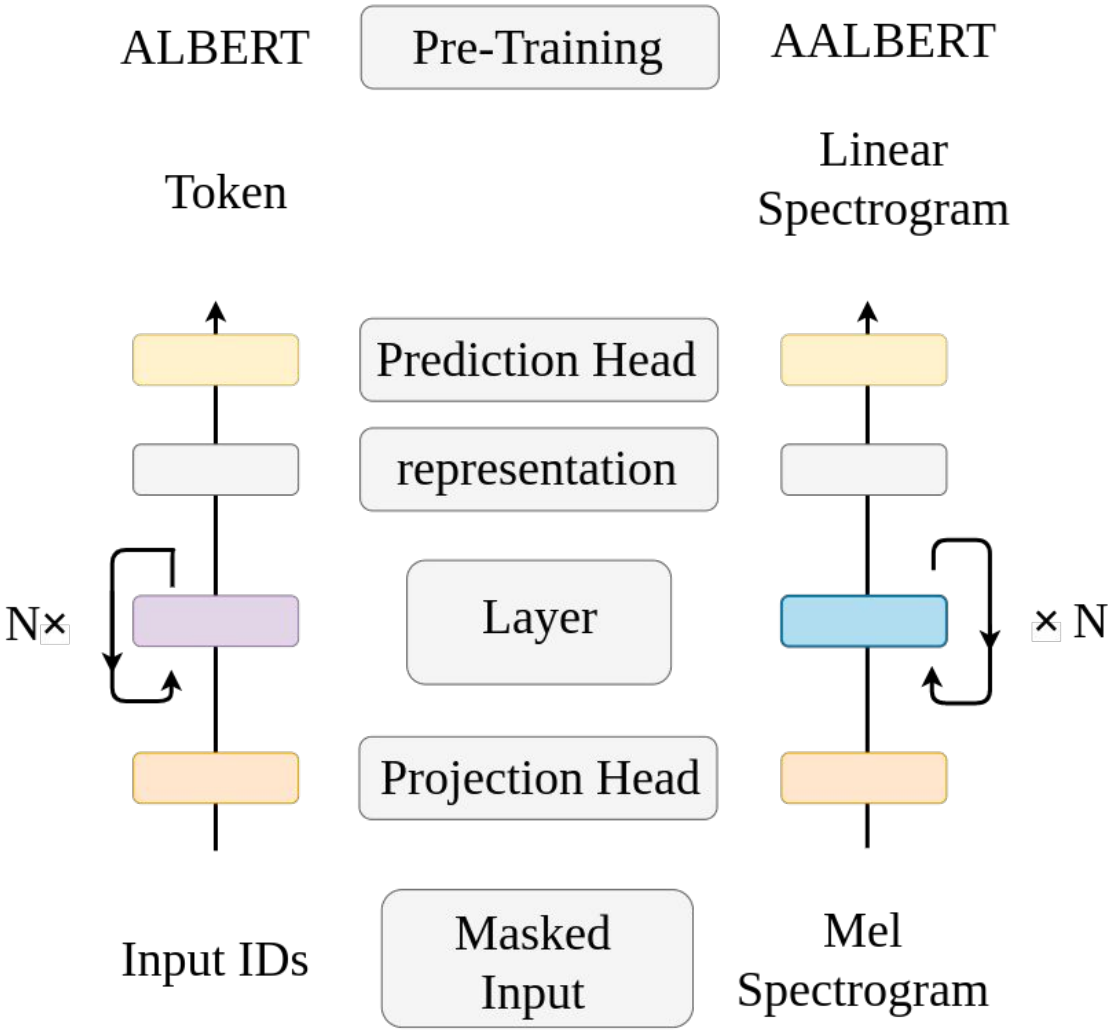
Model Configuration

Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

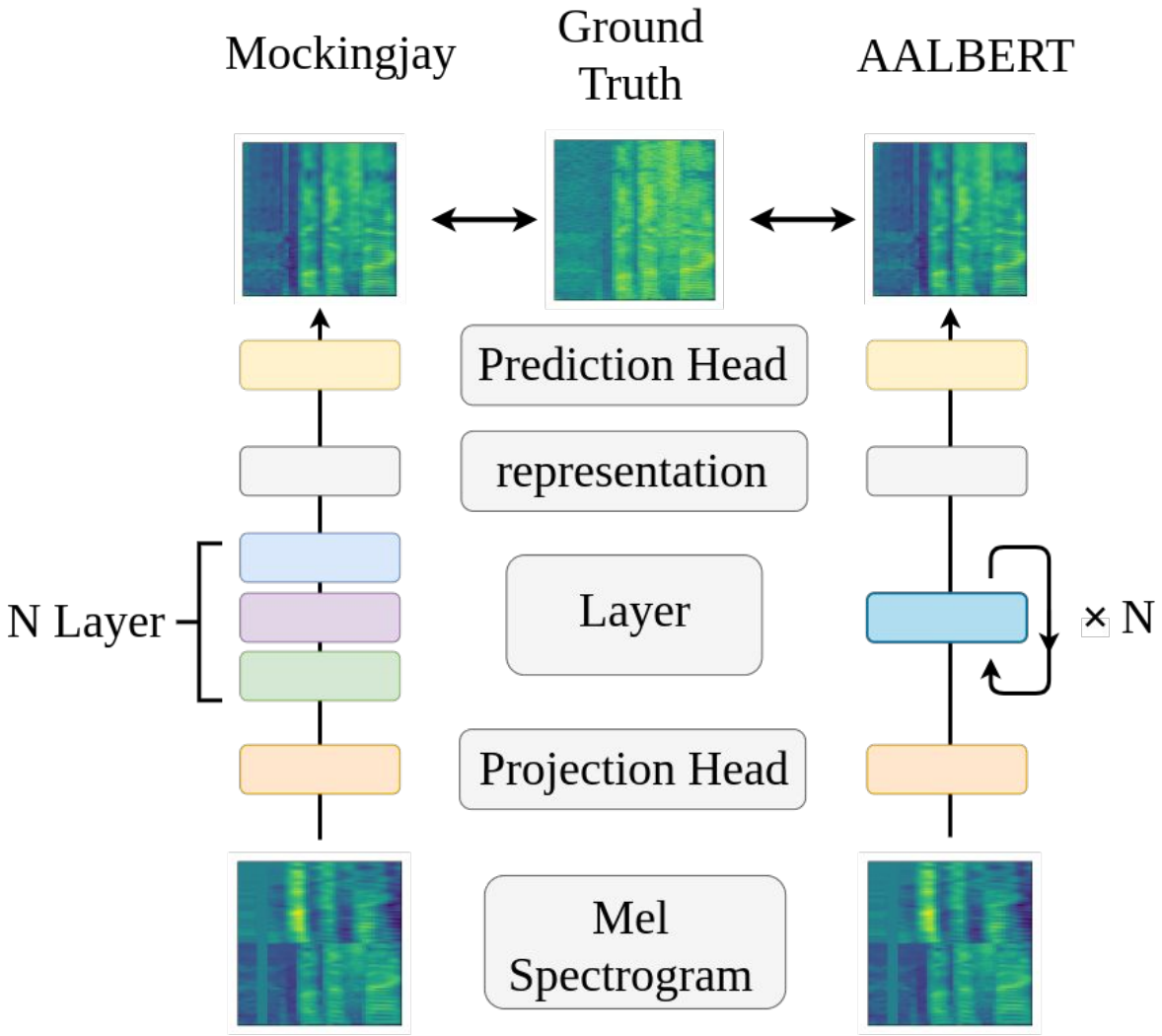
Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg	Speedup
BERT	base	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3	4.7x
	large	334M	92.2/85.5	85.0/82.2	86.6	93.0	73.9	85.2	1.0
ALBERT	base	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1	5.6x
	large	18M	90.6/83.9	82.3/79.4	83.5	91.7	68.5	82.4	1.7x
	xlarge	60M	92.5/86.1	86.1/83.1	86.4	92.4	74.8	85.5	0.6x
	xxlarge	235M	**94.1/88.3**	**88.1/85.1**	**88.0**	**95.2**	**82.3**	**88.7**	0.3x

AALBERT

v.s ALBERT


ALBERT
Pre-Training
AALBERT
Token
Linear
Spectrogram
Prediction Head
representation
N×
Layer
× N
Projection Head
Input IDs
Masked
Input
Mel
Spectrogram

v.s Mockingjay

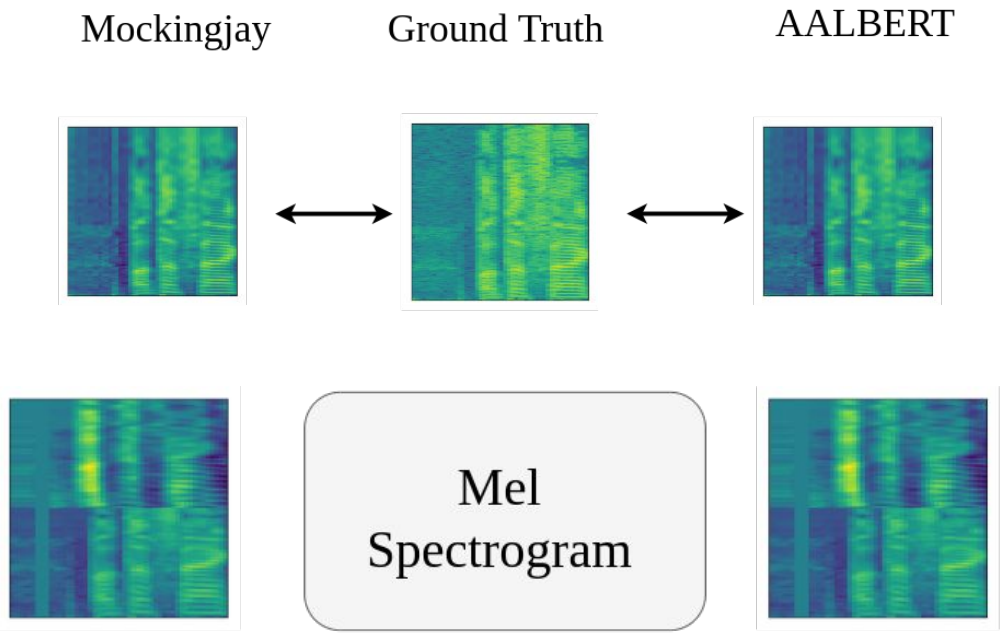


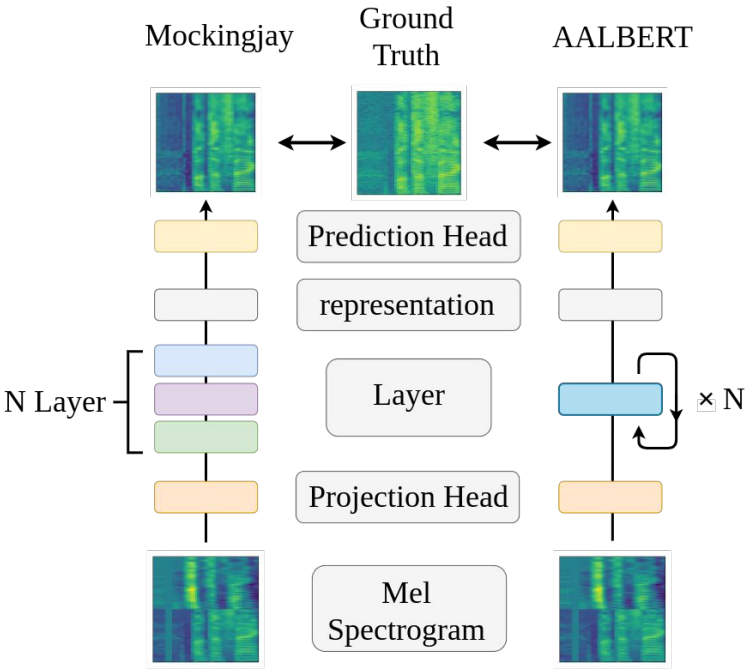
Configuration

Model	Layer	Params	Param Sharing
AALBERT-12L	12	7.4M	True
AALBERT-6L	6	7.4M	True
AALBERT-3L	3	7.4M	True
Mockingjay-12L	12	85.4M	False
Mockingjay-6L	6	42.8M	False
Mockingjay-3L	3	21.4M	False

Pre-Training Stage

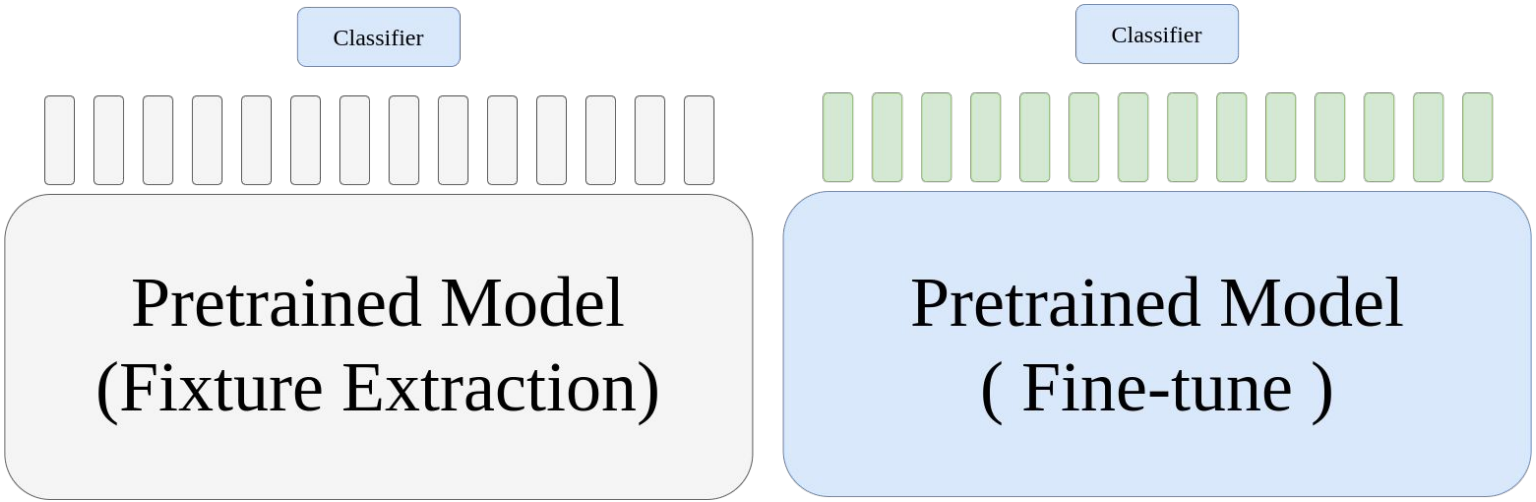
LibriSpeech 360 hours dataset, 500k step, batch size 48.





Phoneme Classification

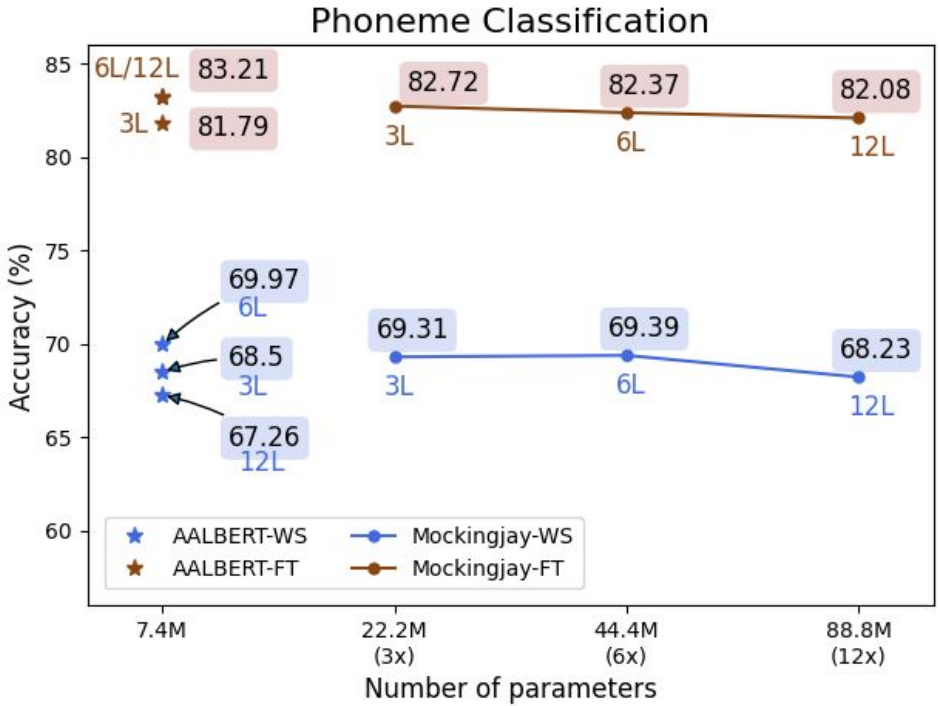
- Weighted-sum and fine-tune feature extraction
- Different Proportion of training data



Phoneme Classification task

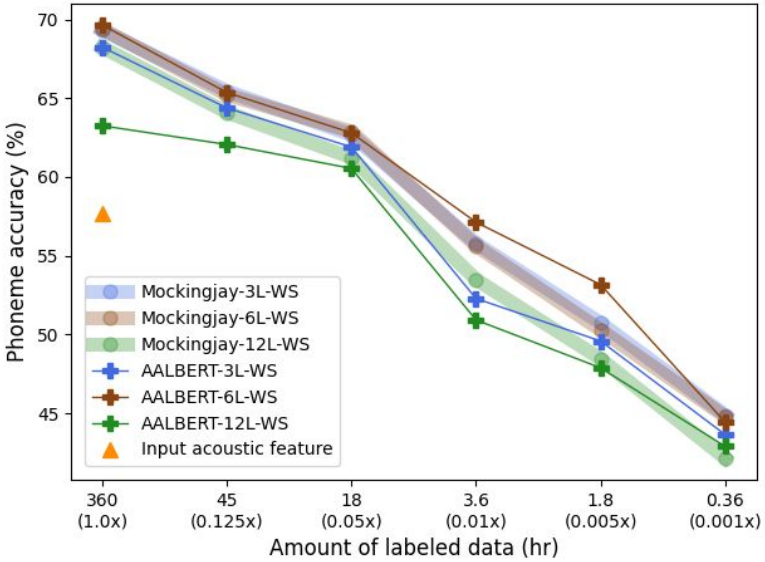
- Utilizing MLP classifier behind representation to train phoneme classification task.
- Weighted-sum, Fine-tune.

Weighted-sum and Fine-tune version

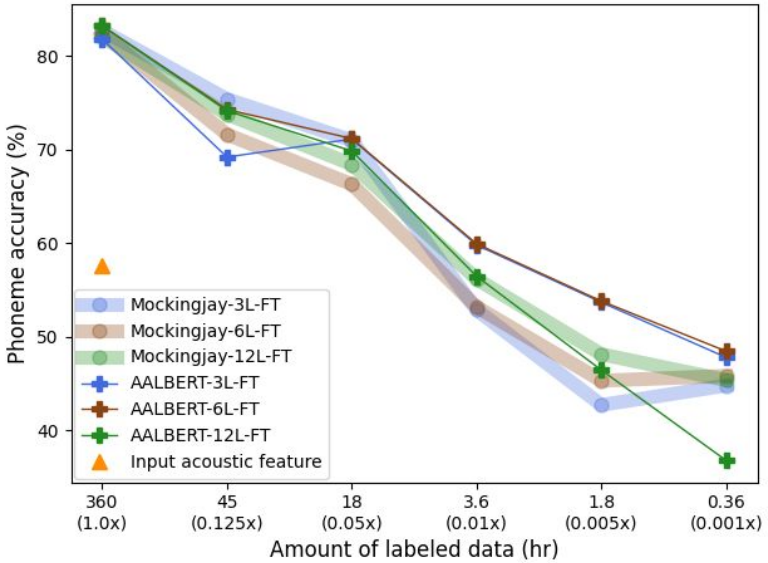


Different Proportion of training data

(Weighted-sum)

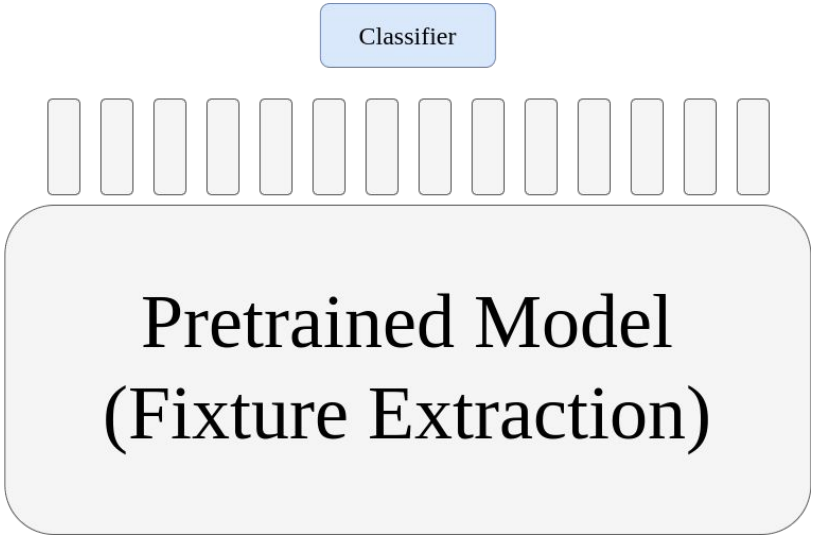


(Fine-tune)



Speaker Identification

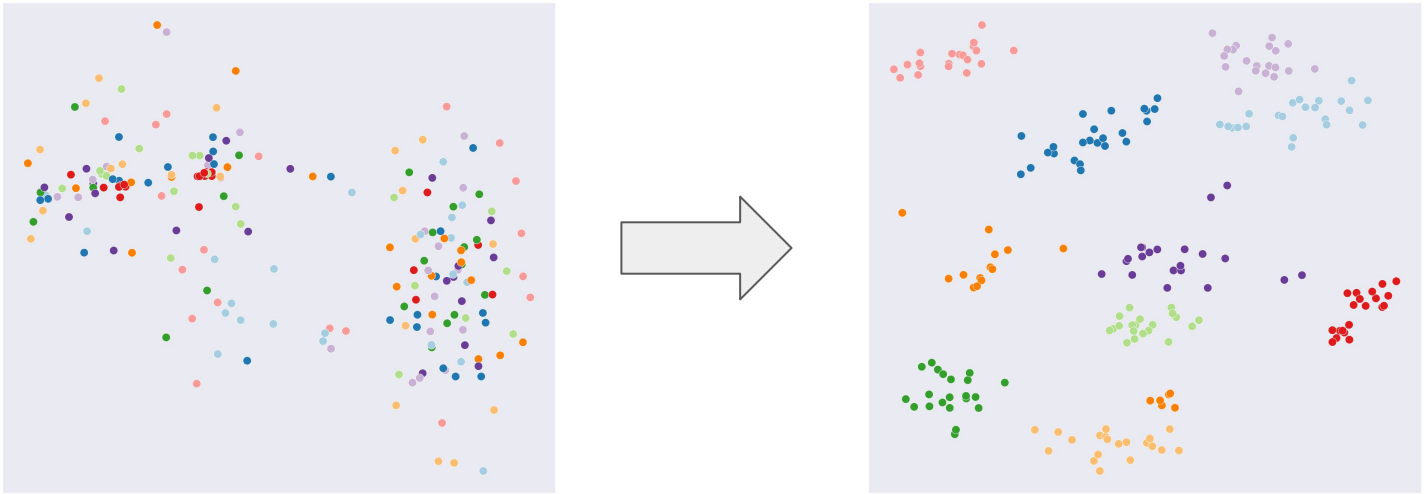
- Utterance-level
- Frame-level
- Overall Performance



Utterance-level

1. Utilizing mean pooling over an utterance to generate utterance-level representation.
2. Simple linear classifier need to train in the Utterance-level speaker identification

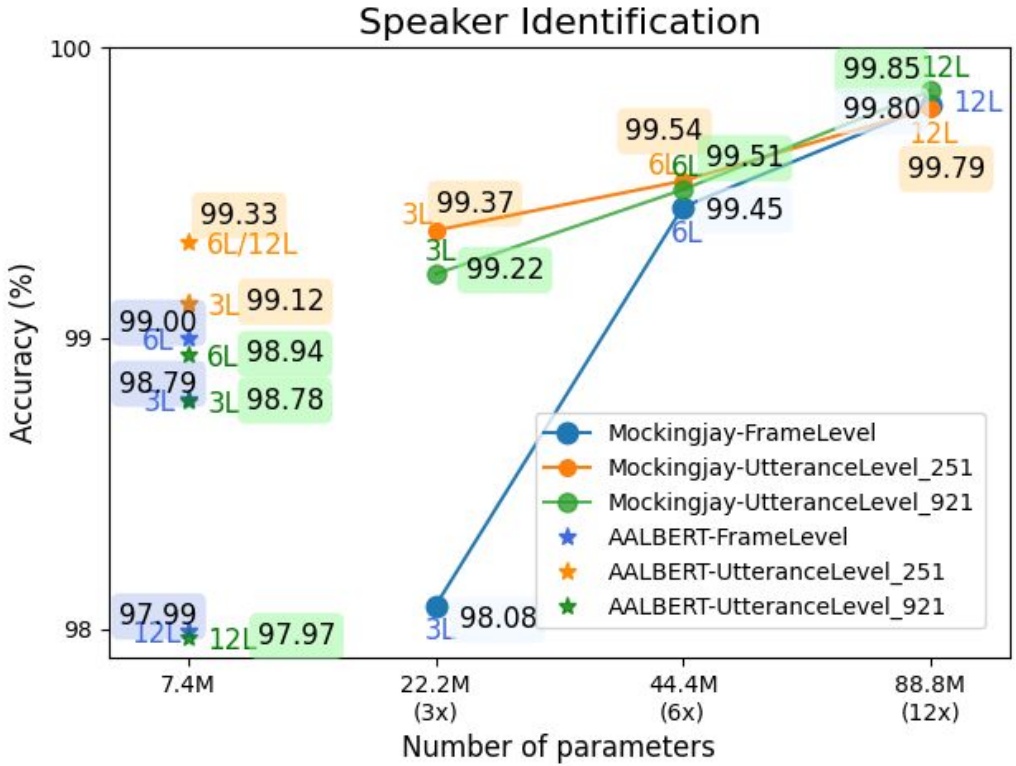
T-sne visualization



Frame-level

1. Classify Each frame-level representation to corresponding speaker.
2. Simple linear Classifier need to train in the frame-level speaker identification

Overall Performance



Probing Tasks