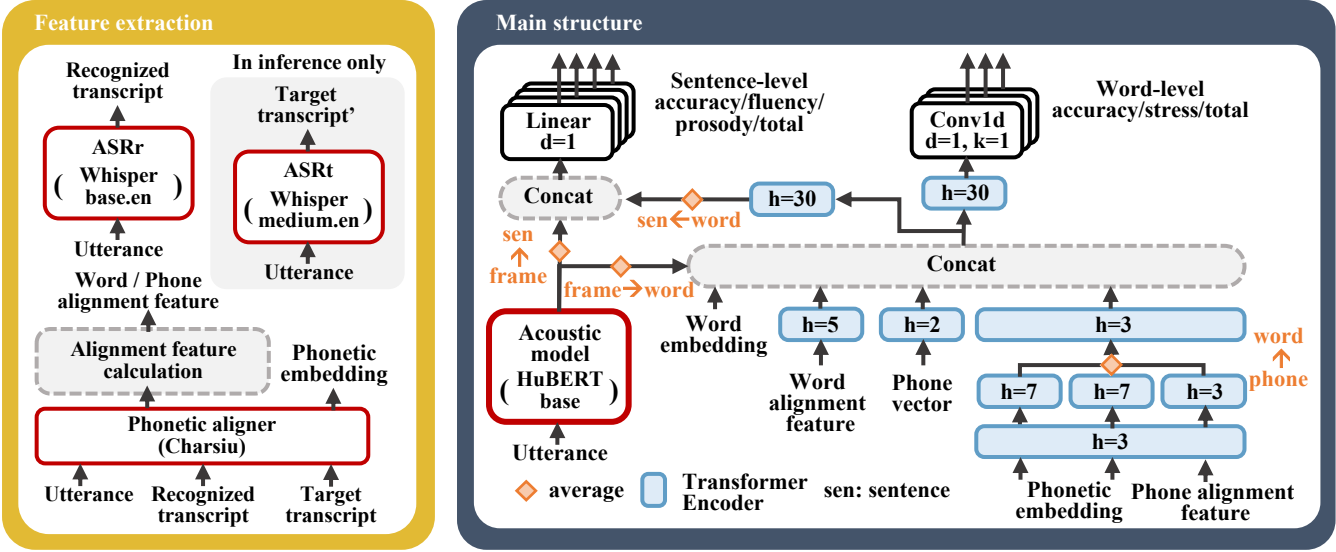
Feature extraction
Recognized transcript
In inference only
Target transcript'
ASRr
(Whisper base.en)
ASRt
(Whisper medium.en)
Utterance
Utterance
Word / Phone alignment feature
Alignment feature calculation
Phonetic embedding
Phonetic aligner (Charsiu)
Utterance
Recognized transcript
Target transcript
Main structure
Linear d=1
Sentence-level accuracy/fluency/ prosody/total
Conv1d d=1, k=1
Word-level accuracy/stress/total
Concat
sen←word
h=30
h=30
sen
frame
frame→word
Concat
Acoustic model (HuBERT base)
Word embedding
h=5
h=2
h=3
word
phone
Word alignment feature
Phone vector
h=7
h=7
h=3
h=3
Utterance
average
Transformer Encoder
sen: sentence
Phonetic embedding
Phone alignment feature