Tokenwise Contrastive Speech and Text pre-training for Emotion Recognition

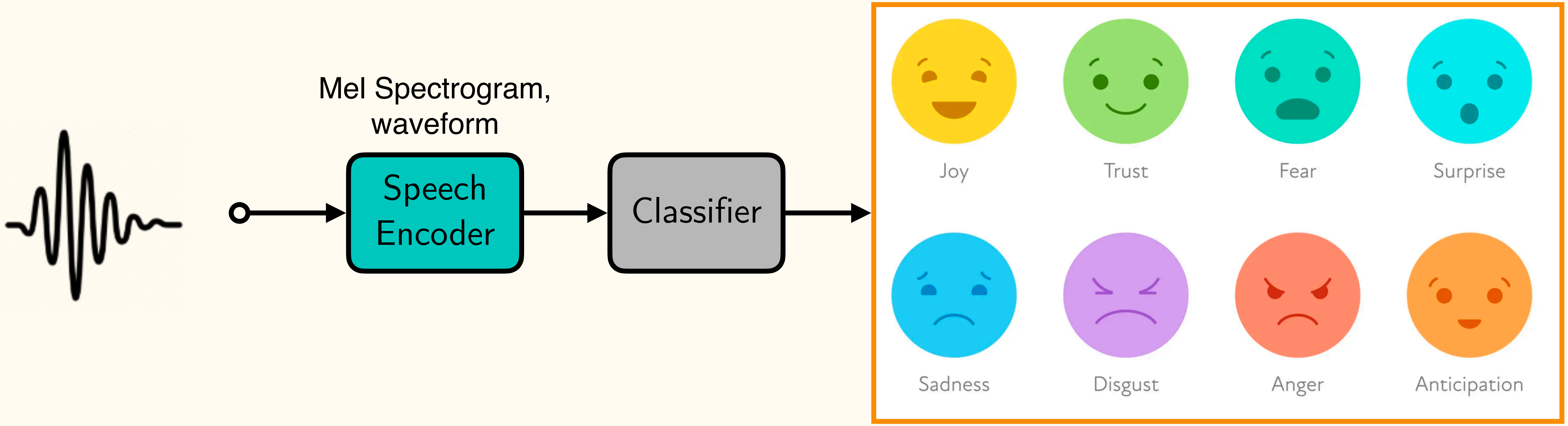
Eklavya SARKAR and Neha TARIGOPULA

EE608 - Deep Learning for Natural Language Processing

Table of Contents

Speech Emotion Recognition (SER)

- Recognize human emotion and affective states from oral speech.
- Essential task in the human-computer interaction (HCI) field.
- Useful in applications such as call-center bots or intelligent cars.

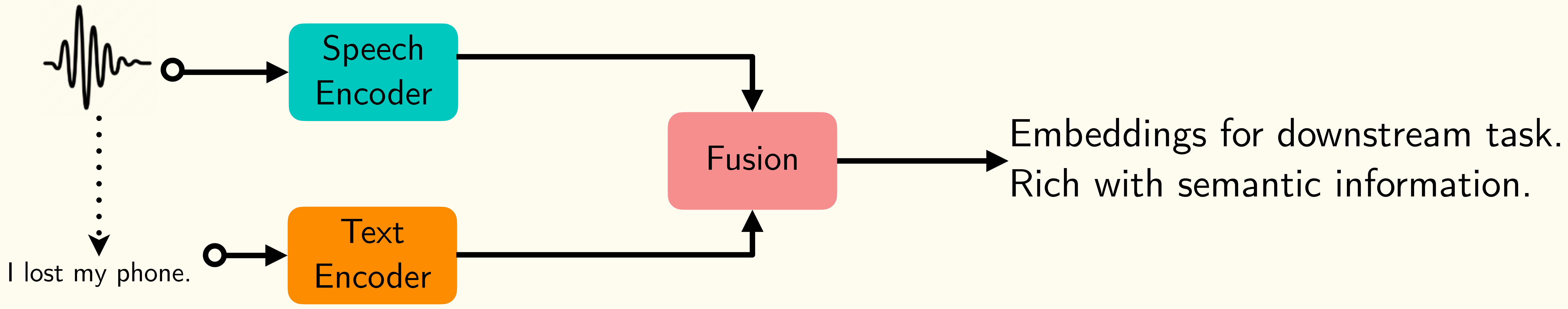


Motivation

- Common approaches would typically use audio features.
  - However, these features focus only on paralinguistic acoustic/spectral information.
  - No information on semantic (language) knowledge from the spoken words.

Research Question:

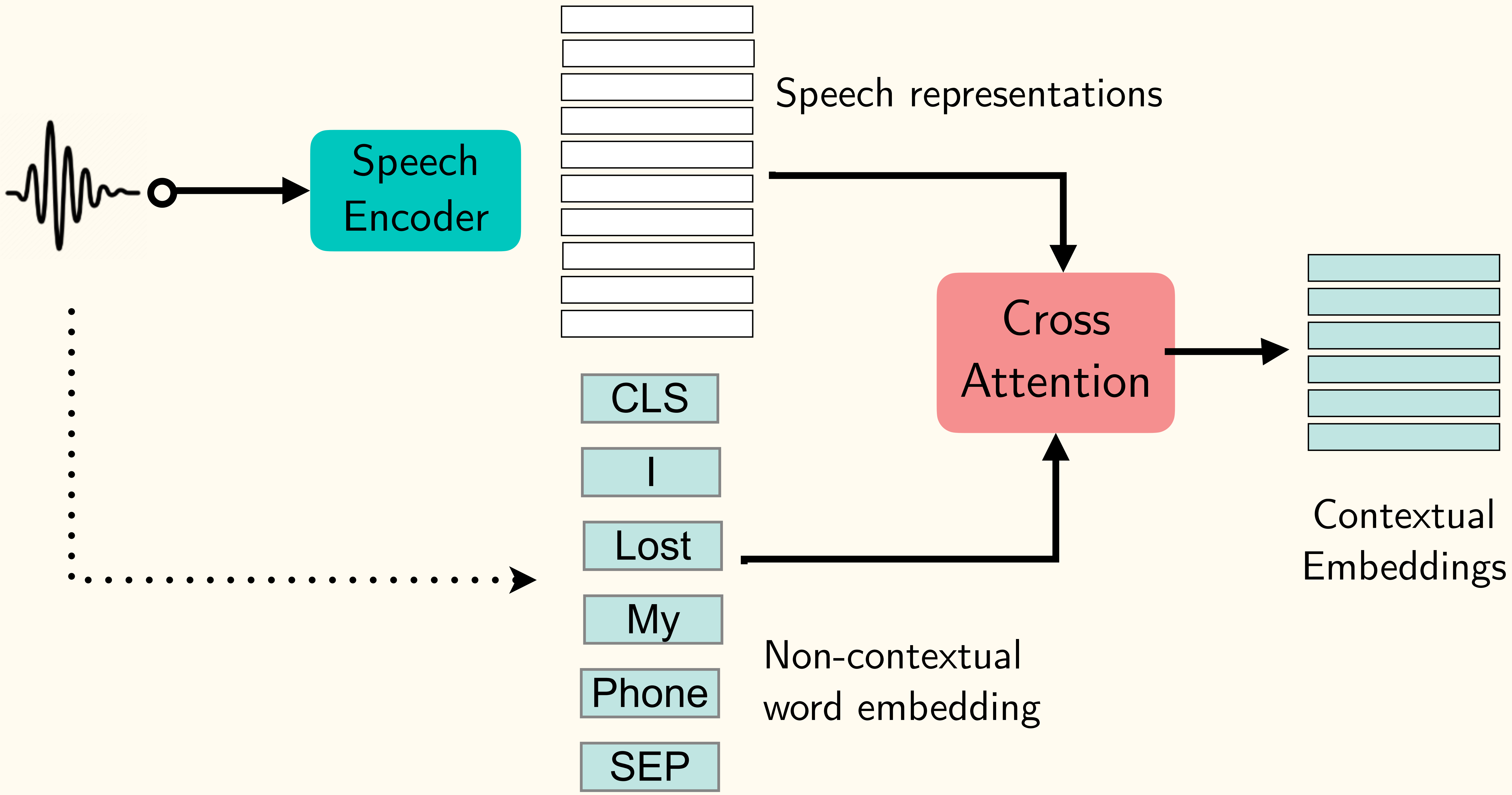
- Can leveraging additional textual information improve representations for speech emotion recognition ? (Untapped potential)



Proposed Method

Distill knowledge from BERT to audio embeddings via token-by-token alignment of speech and text

1. Use the speech representation of an utterance to convert a **non-contextual** word embedding (of the corresponding utterance's transcript) → *to* **contextual** embedding tokens by using a *cross-modal attention mechanism*.



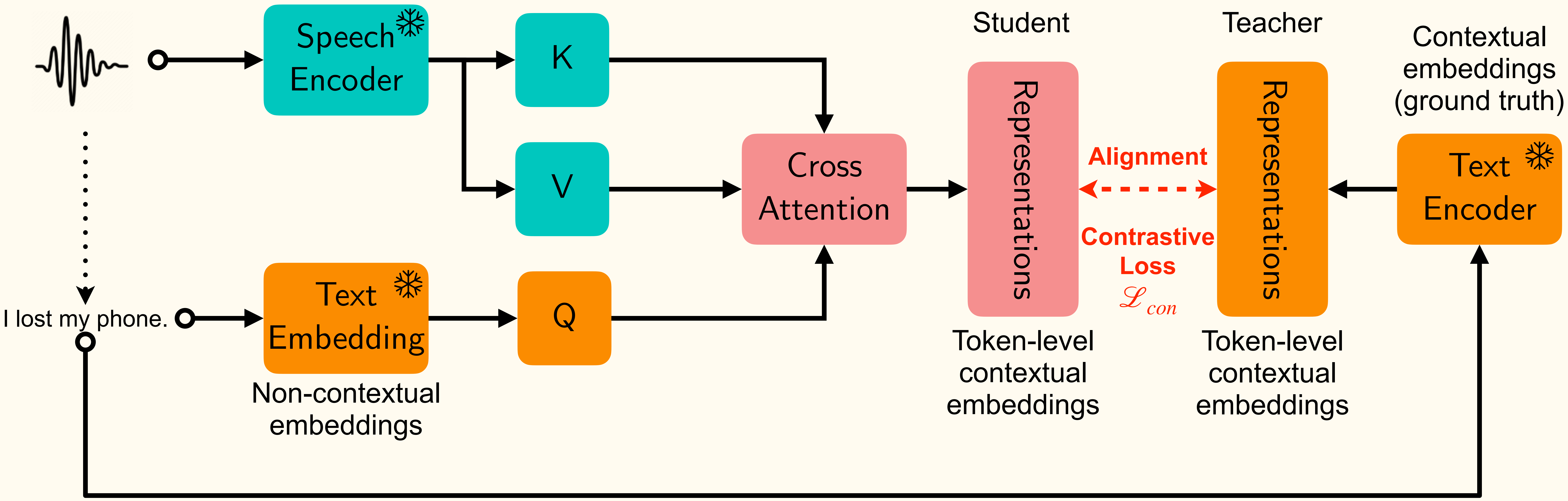
Proposed Method

1. Use the speech representation of an utterance to convert a **non-contextual** word embedding (of the corresponding utterance's transcript) → *to* **contextual** word embeddings by using a *cross-modal attention mechanism*.
2. Use a contrastive loss to implicitly inject fine-grained semantic knowledge from a 'ground truth' (contextualized) text-encoder into the speech representations.

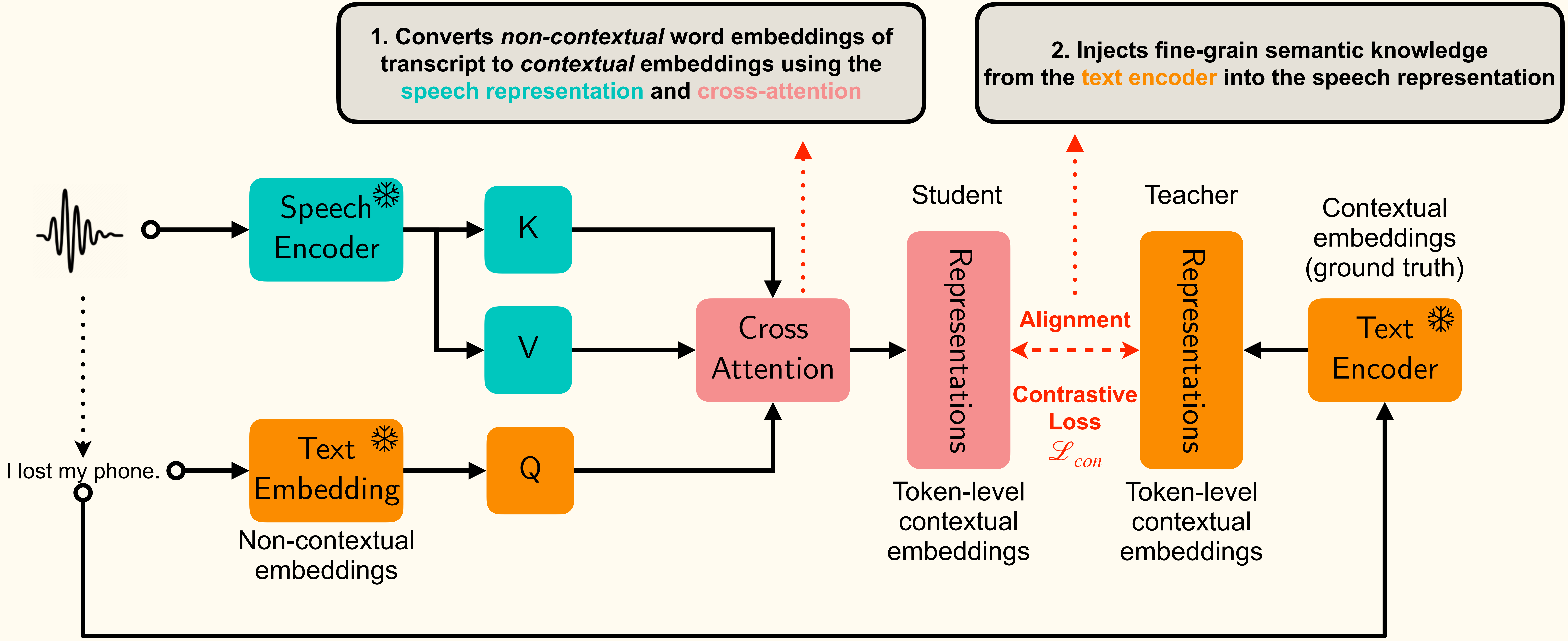
- Previous work[1] has shown results for speech2intent tasks
  - Hasn't been tested on SER systems.

[1] *Tokenwise Contrastive Pretraining for Finer Speech-to-BERT Alignment in End-to-End Speech-to-Intent Systems* (2022), Sunder et al., Interspeech.

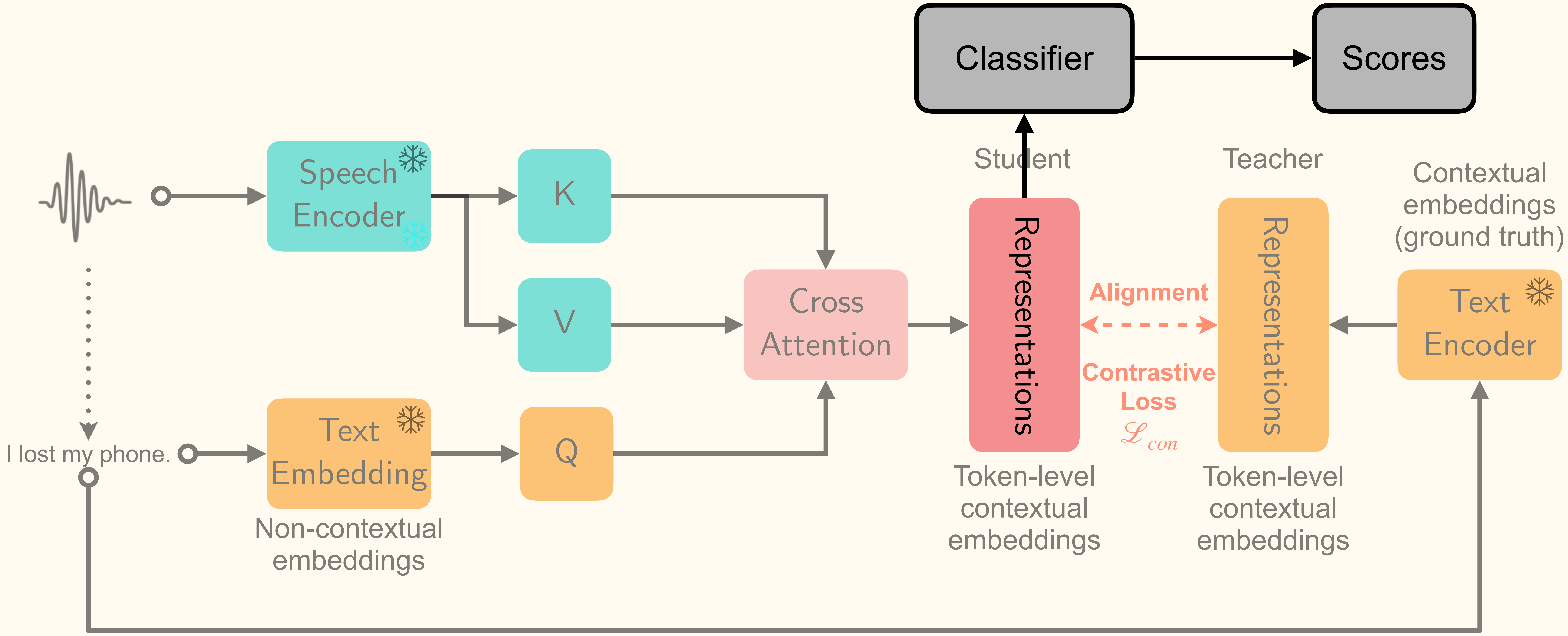
Experiment Design - Pretraining



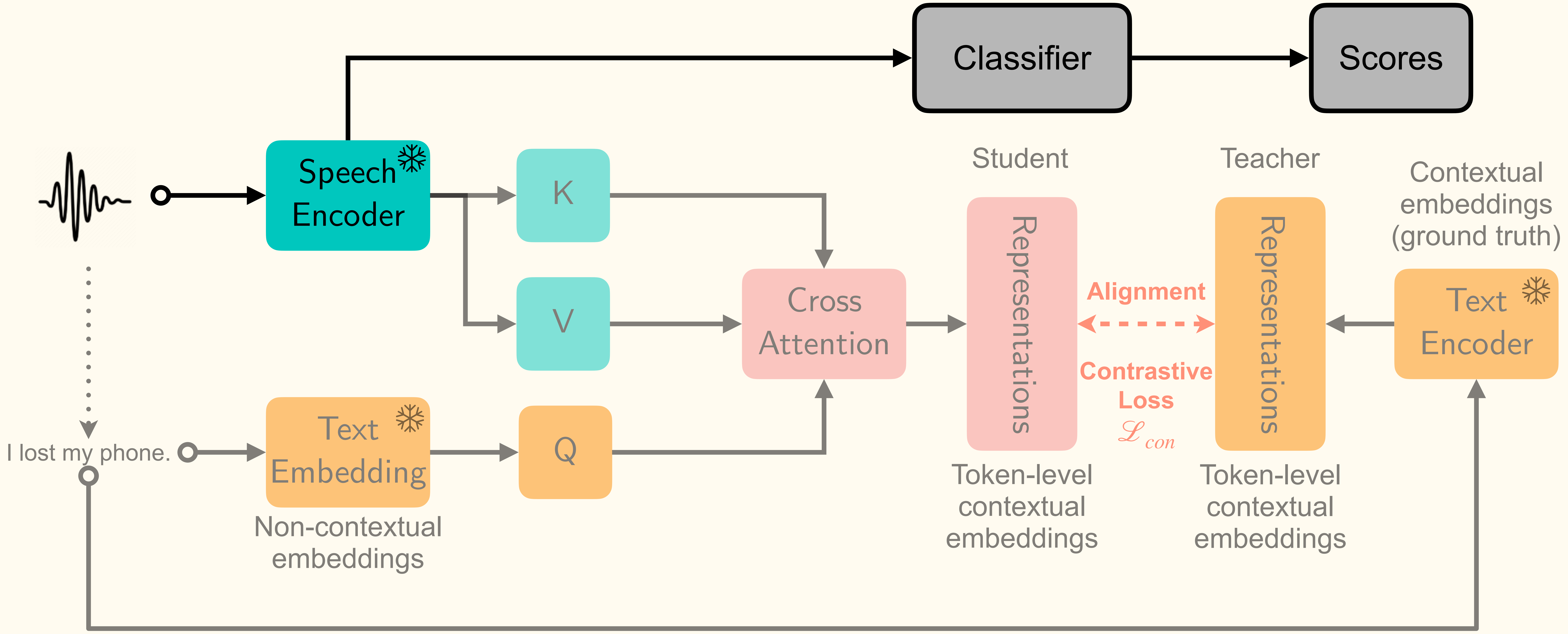
Experiment Design - Pretraining



Experiment Design - Downstream



Experiment Design - Baseline



Experiment Setup

Baseline (*speech* only):

- Whisper embeddings (last encoder layer) $\in \mathbb{R}^{N \times D}$.
- Compute and concatenate statistics (μ, σ) into functional vector $\in \mathbb{R}^{2D}$.

Proposed (*speech x text*):

- Input features: Whisper x BERT contextualized embeddings.

Classifier:

- Simple feedforward network.
- 3 x [Linear, LayerNorm, ReLUs].

Metrics:

- Accuracy and F1-score.

Protocols:

- 70:20:10 split into *Train*, *Val*, *Test*.

Experiment Setup

Downstream Datasets:

- EmoDB:
  - 7 classes.
- IEMOCAP:
  - 5 classes.

- 1 utterance = 1 emotion.
- Recorded by 10 actors (5 male, 5 female).
- Scripted and improvised.
- 16 kHz.

Utterances and labels per dataset.

Emotion	**EmoDB**	**IEMOCAP**
Ang	127	1103
Hap	71	595
Neu	79	1708
Sad	62	1084
Dis	46	-
Fea	69	-
Bor	81	-
Exc	-	1041
Total	**535**	**5531**

Experiment Setup

Downstream Datasets:

- EmoDB:
  - 7 classes.
- IEMOCAP:
  - 5 classes.

- 1 utterance = 1 emotion.
- Recorded by 10 actors (5 male, 5 female).
- Scripted and improvised.
- 16 kHz.

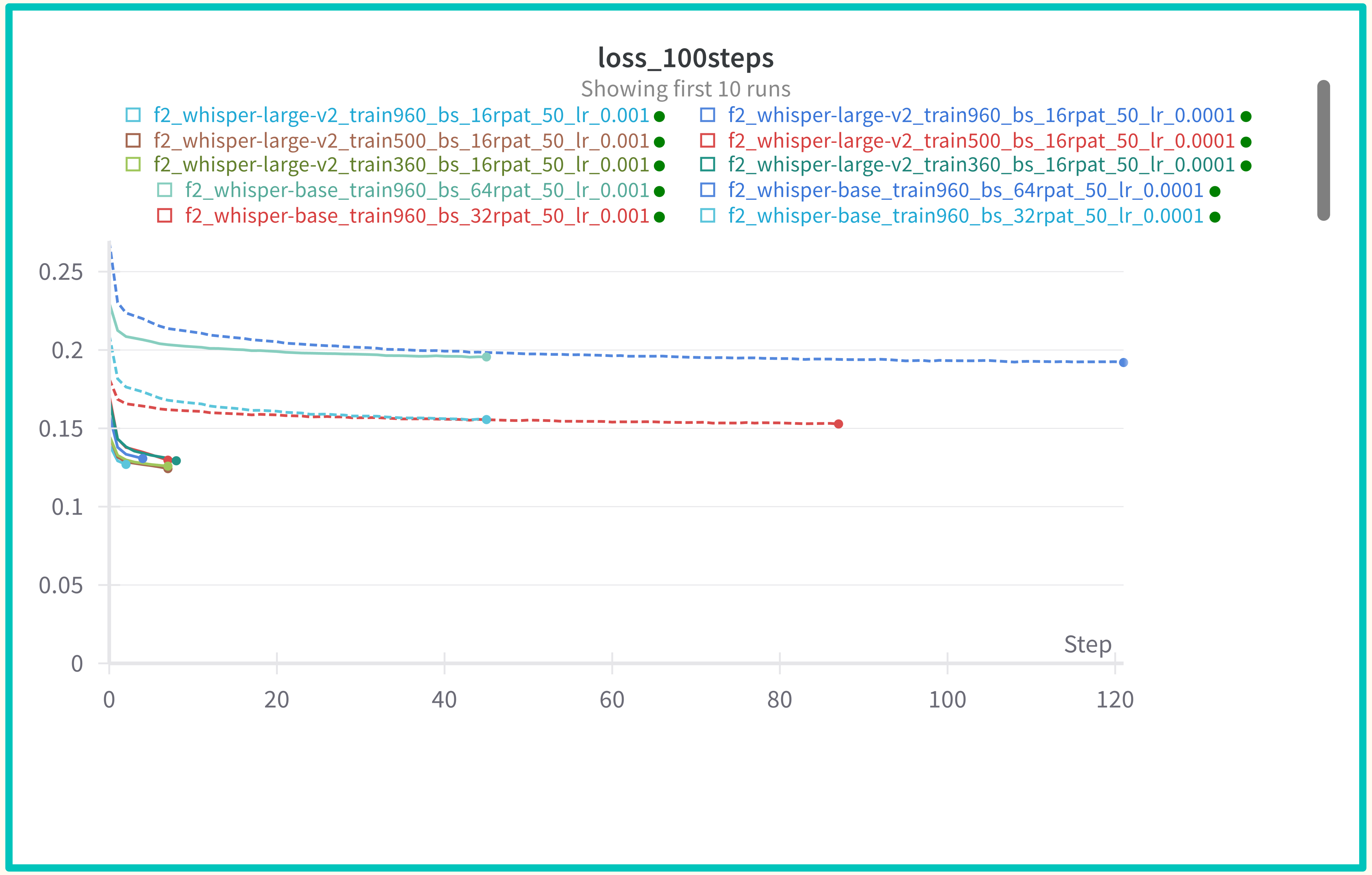
Pre-training Dataset:

- Librispeech:
  - Train 100h
  - Train 500h
  - Train 960h

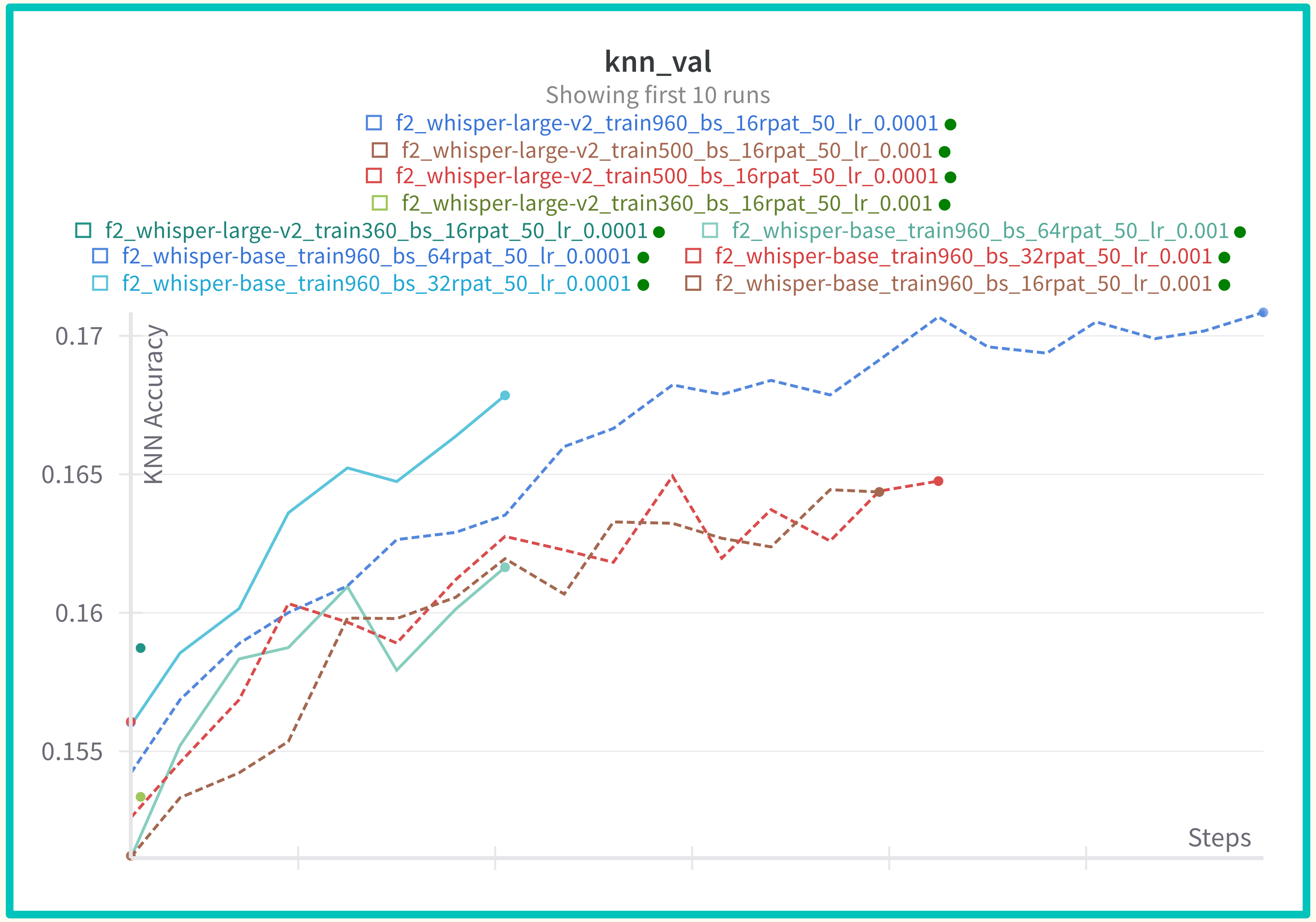
  - Read English (audiobooks).
  - 16 kHz.

Ongoing Work - Pretraining

Pre-training loss



Validation



Ongoing Work - Baselines

Accuracy and F1 scores [%] on *Test*.

Method	**DB**	**Acc**	**F1**
Baseline	EmoDB	76.2	81.5
	IEMOCAP	–	–
Proposed	EmoDB	–	–
	IEMOCAP	–	–

Search space to find optimal hyper-parameters.

Classifier	**Hyperparams**	**Search space**
Baseline	Batch size	2**[2, 10]
	Learning rate	1e{-3, -2}
Proposed	Batch size	2**[2, 10]
	Learning rate	1e{-3, -2}
	Model	Base, Large

Summary and Future Work

- Results will show whether this cross-alignment will help for SER.

Future Work:

- Try other encoders:
  - Text: word2vec, GloVe.
  - Audio: WavLM.

Thank you!