Prompting Audios Using Acoustic Properties for Emotion Representation

Hira Dhamyal, Benjamin Elizalde, Soham Deshmukh, Huaming Wang, Bhiksha Raj, Rita Singh

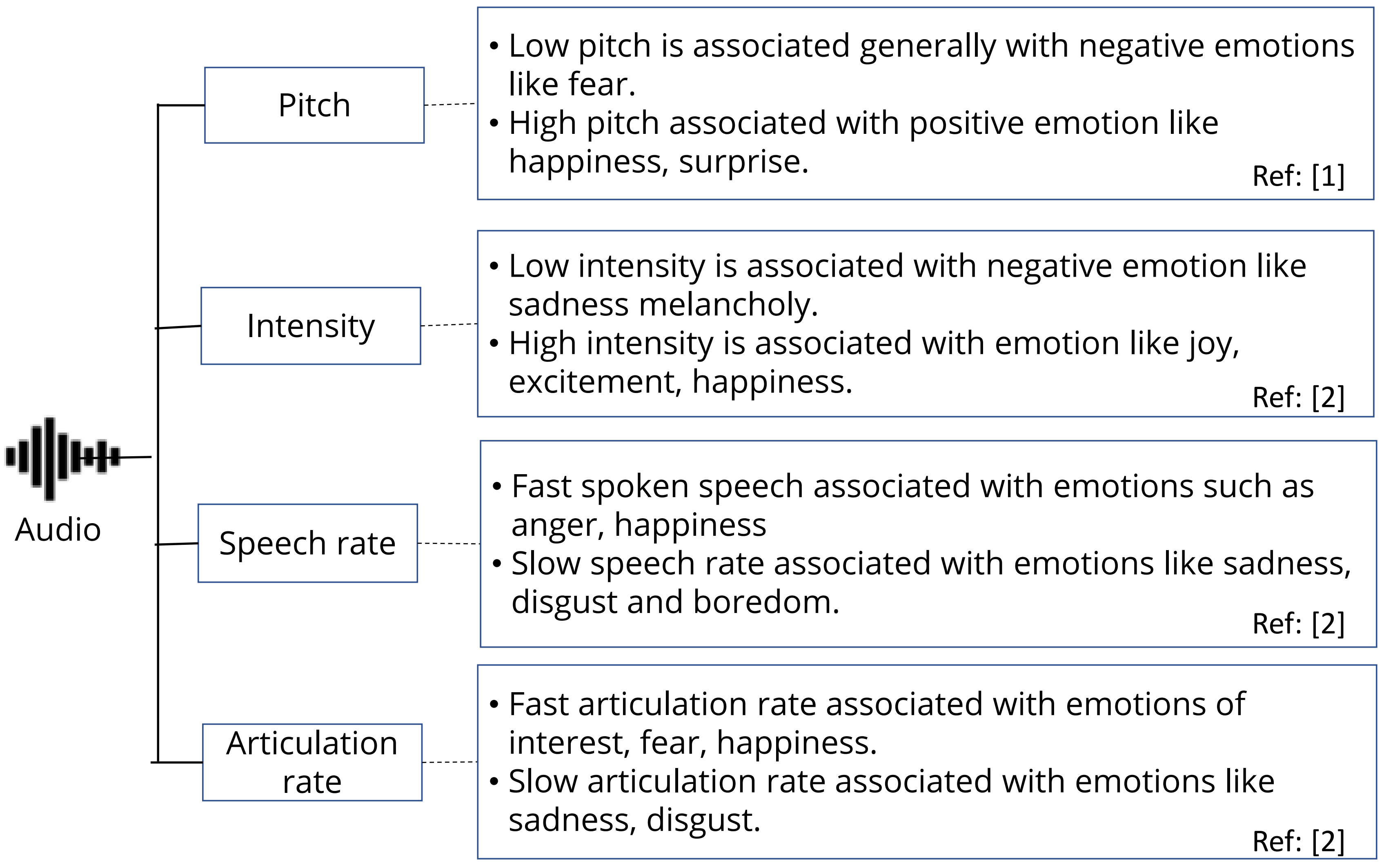
Introduction

- Emotions lie on a continuum, but current models treat emotions as a finite valued discrete variable. This representation does not capture the diversity in the expression of emotion. **To better represent them, we propose the use of natural language descriptions (or prompts).**
- In this work, we address the challenge **of automatically generating these prompts** and training a model to better learn emotion representations from **audio and prompt pairs.**
- We use **acoustic properties** that are correlated to emotion like pitch, intensity, speech rate, and articulation rate to automatically generate prompts, i.e., **'acoustic prompts'**.

DATASETS

Dataset	# Files	# Classes	Emotions
CMU-MOSEI	23K	9	ang, exc, fear, sad, frus, neu, sur, hap, dis
IEMOCAP	10K	9	hap, fear, sad, sur, exc, ang, neu, disappoint, frus
MELD	10K	7	neu, sur, fear, sad, joy, disgust, ang
CREMA-D	7K	6	ang, dis, fear, hap, neu, sad
RAVDESS	2.5K	8	neu, calm, hap, sad, ang, fear, disgust, sur
CMU-MOSI	2.2K	3	neu, positive, negative

MOTIVATION



- Low pitch is associated generally with negative emotions like fear.
- High pitch associated with positive emotion like happiness, surprise.

Ref: [1]

- Low intensity is associated with negative emotion like sadness melancholy.
- High intensity is associated with emotion like joy, excitement, happiness.

Ref: [2]

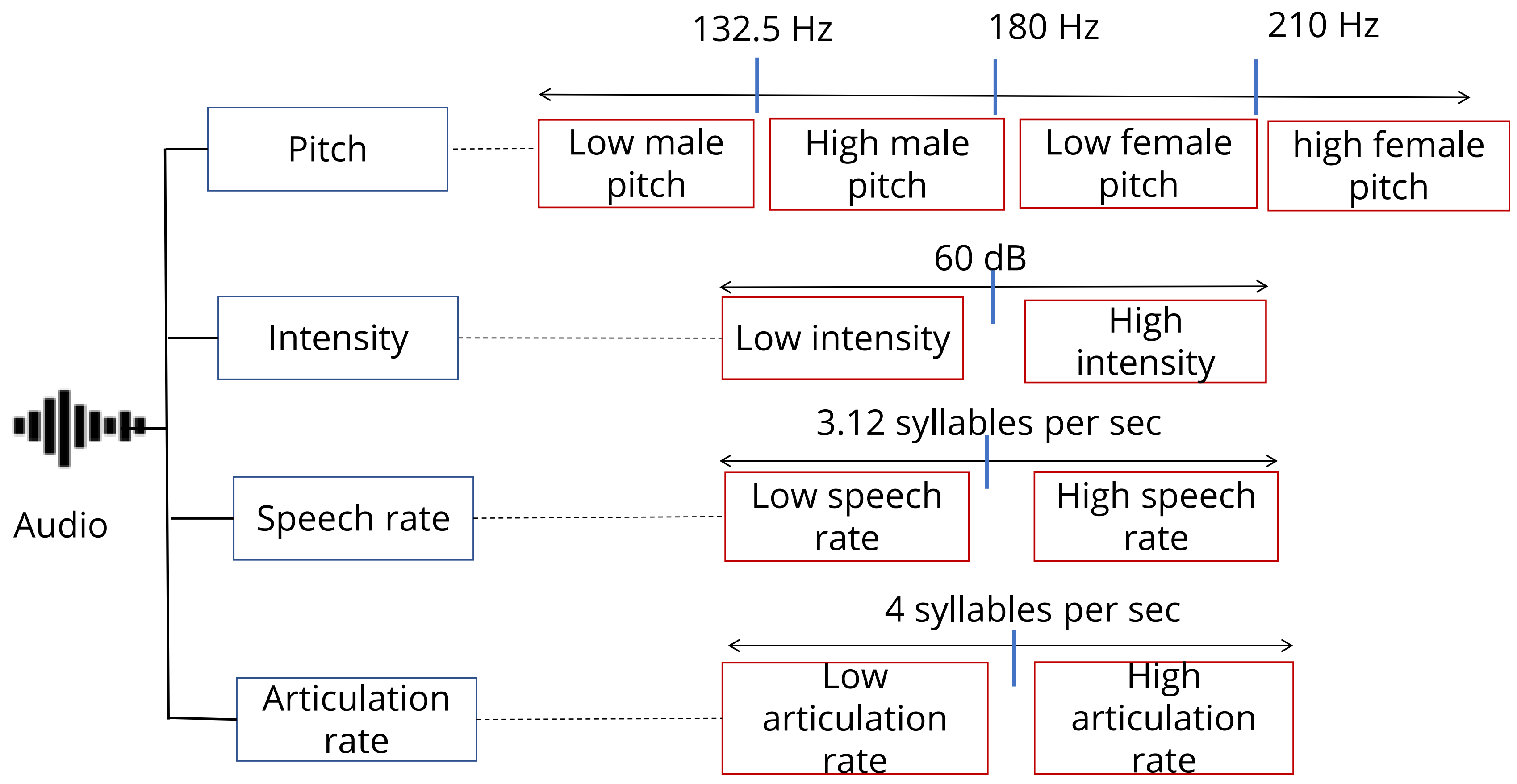
- Fast spoken speech associated with emotions such as anger, happiness
- Slow speech rate associated with emotions like sadness, disgust and boredom.

Ref: [2]

- Fast articulation rate associated with emotions of interest, fear, happiness.
- Slow articulation rate associated with emotions like sadness, disgust.

Ref: [2]

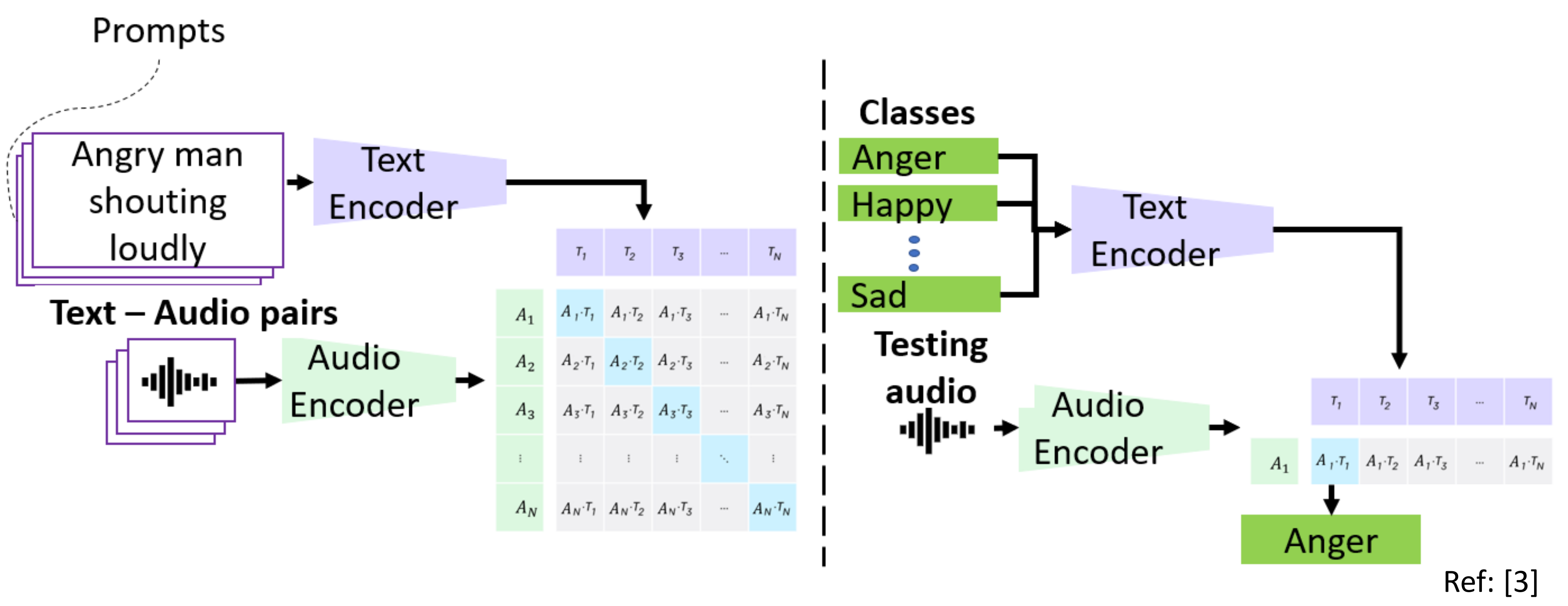
METHODOLOGY



EXAMPLE PROMPTS

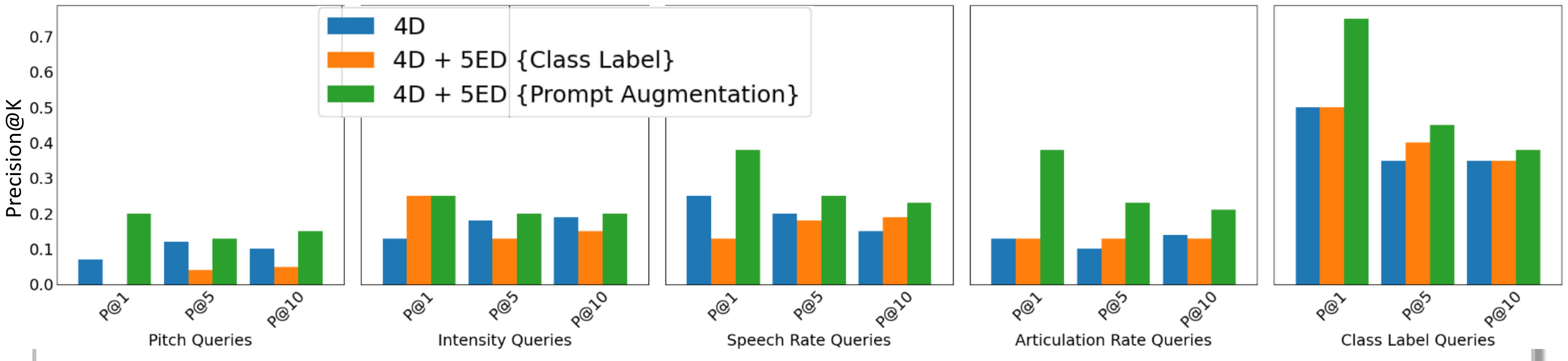
Property	Prompt
Class Label	• {emotion}
Pitch	• High female pitch {emotion} • High male pitch {emotion}
Intensity	• High intensity {emotion}
Speech Rate	• High speech rate {emotion}
Articulation Rate	• High articulation rate {emotion}

MODEL


Prompts
Angry man shouting loudly
Text Encoder
Text – Audio pairs
Audio Encoder
T_1 T_2 T_3 … T_N
A_1 A_2 A_3 A_N
$A_1{\cdot}T_1$ $A_1{\cdot}T_2$ $A_1{\cdot}T_3$ … $A_1{\cdot}T_N$
$A_N{\cdot}T_1$ $A_N{\cdot}T_2$ $A_N{\cdot}T_3$ … $A_N{\cdot}T_N$
Classes
Anger
Happy
Sad
Text Encoder
Testing audio
Audio Encoder
Anger

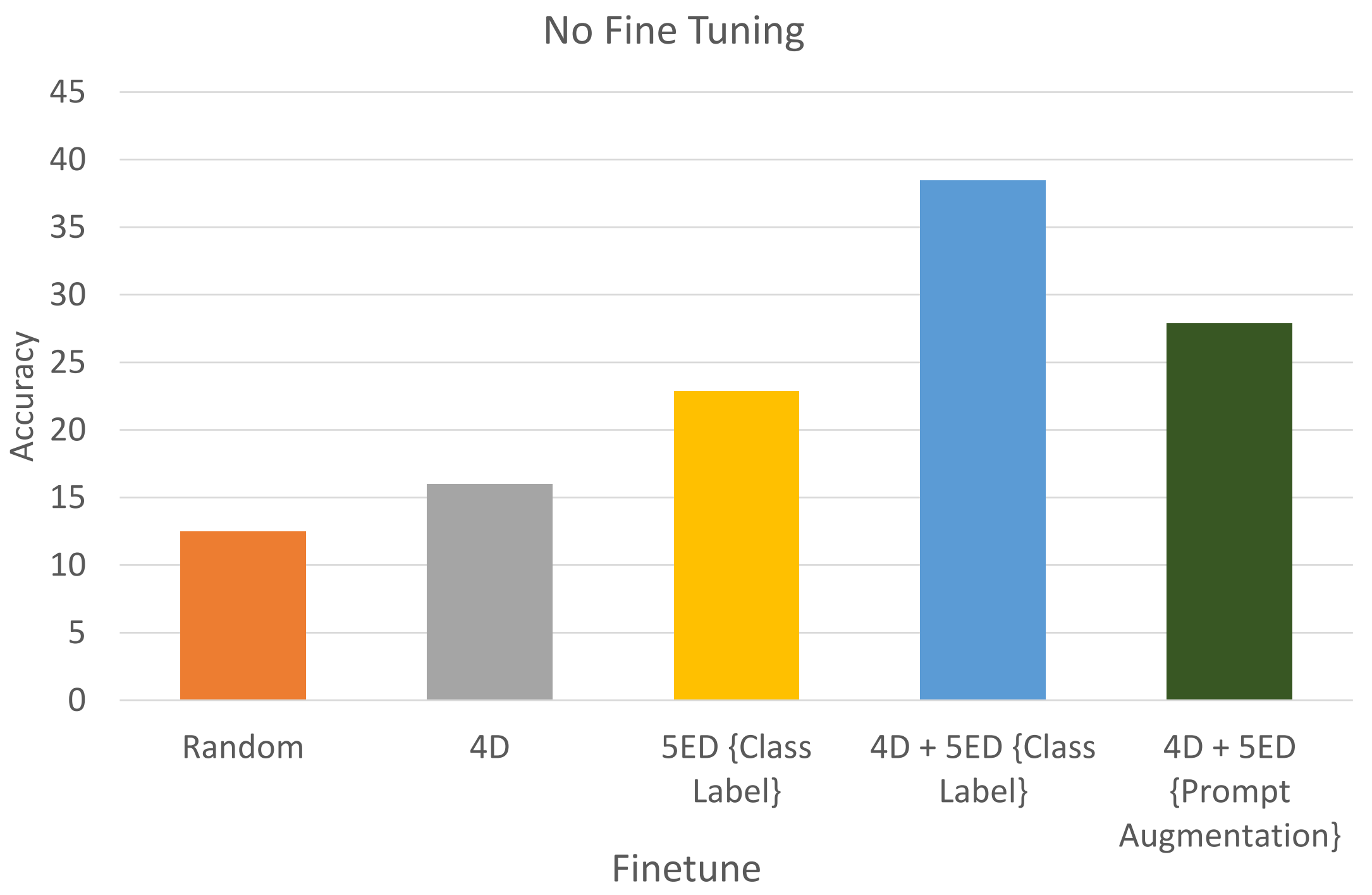
Ref: [3]

RESULTS – Emotion Audio Retrieval

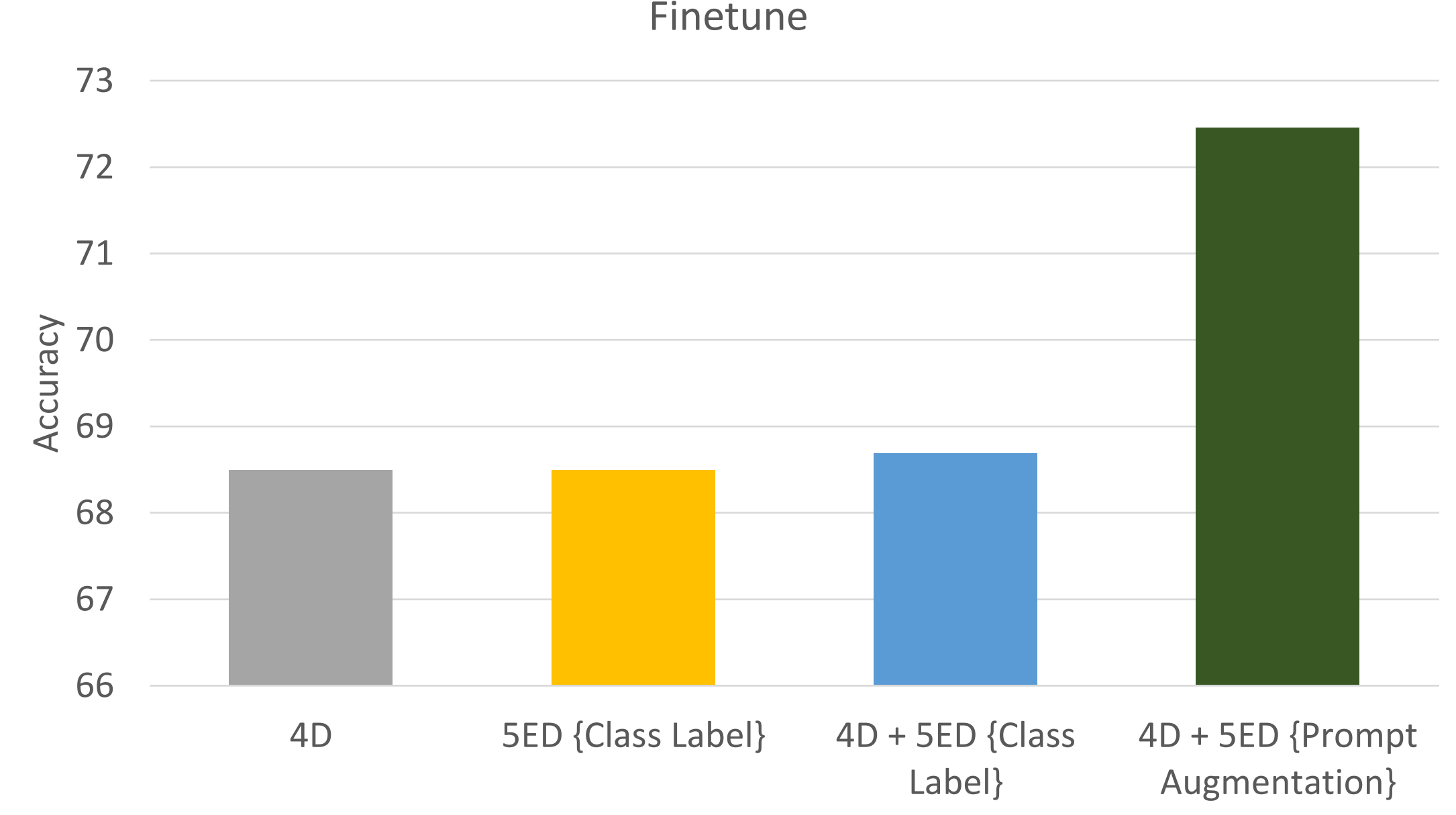


RESULTS – Speech Emotion Recognition

- The performance is shown on RAVDESS dataset. The dataset is not included in the training datasets.

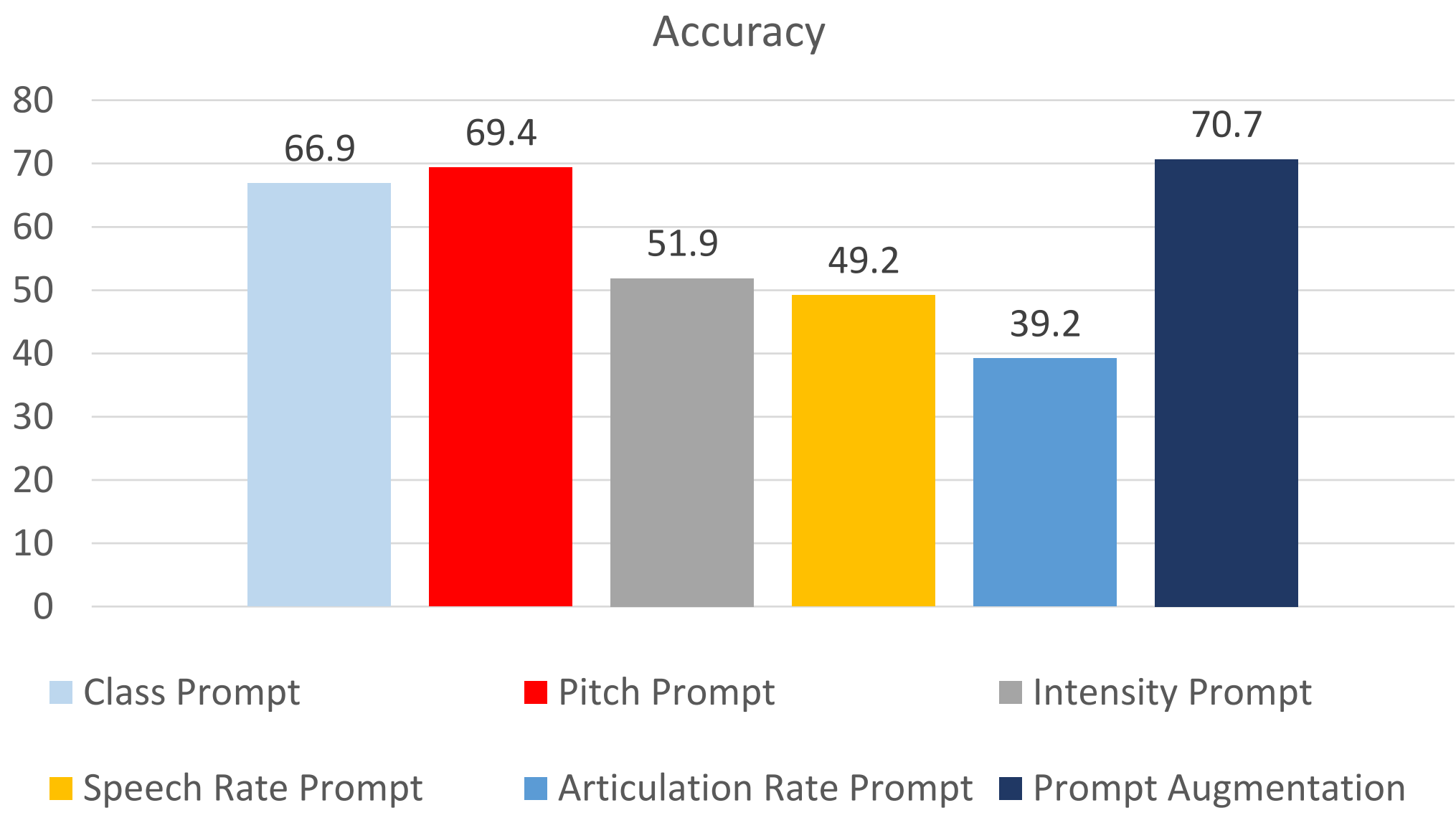


- The model is finetuned on the training subset of RAVDESS and performance shown on testing subset.



ANALYSIS – which Prompt is better?

- Performance shown when the model is finetuned on RAVDESS dataset using different acoustic prompts paired with the audio.



CONCLUSION

- We find that among the acoustic prompts, pitch prompt is the best performing one.
- Emotion Audio Retrieval - acoustic prompt augmentation achieves consistently better Precision@K metric.
- Speech Emotion Recognition - shows performance improvement by 3.8% absolute in RAVDESS.

[1] Scherer, Klaus R. Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences. (1972).
[2] Pavlenko, Aneta. Emotions and multilingualism. Cambridge University Press, 2005.
[3] Benjamin Elizalde, et al "Clap: Learning audio concepts from natural language supervision," arXiv preprint arXiv:2206.04769, 2022.