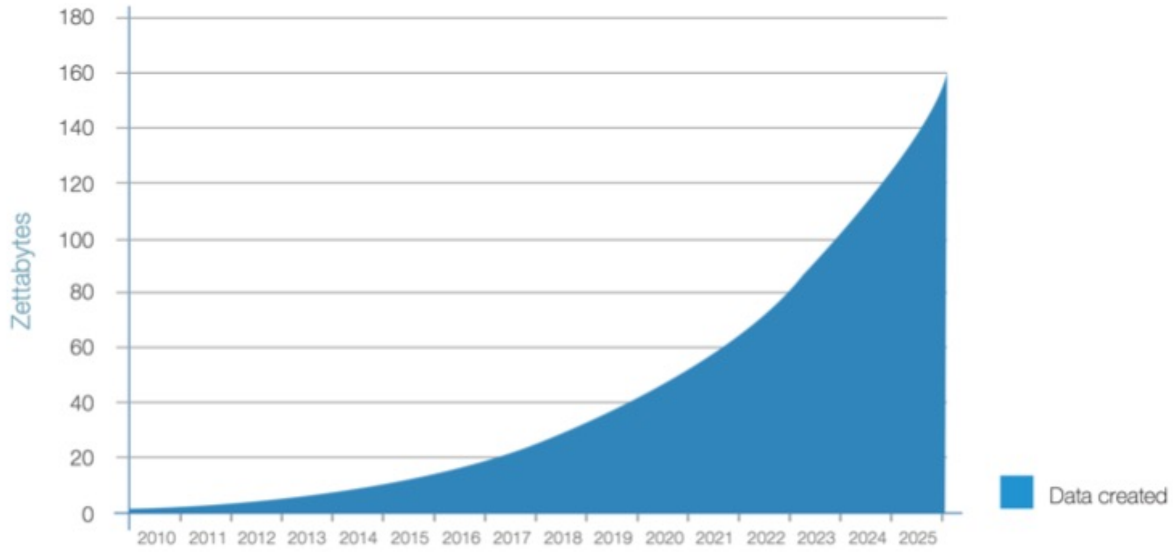
Self-supervised Speaker Recognition Training Using Human-Machine Dialogues

Metehan Cekic[1*], Ruirui Li[2], Zeya Chen[2], Yuguang Yang[2], Andreas Stolcke[2], Upamanyu Madhow[1]

[1]University of California, Santa Barbara
[2]Amazon Alexa AI

*Work done during an internship at Amazon

Background and Motivation



A. Source: IDC DataAge 2025 whitepaper

- Spread of smart devices → **Exponential data growth**
- Most of the collected data is **without labels**.
- **Labeling data** is **cumbersome** and **expensive**.

Background and Motivation

- **Do we really need labels?**
  - How can a neural network learn representations without labels?
- **A promising approach: Contrastive Learning**
  - Learn the **general features of a dataset** by teaching the model which **data points** are **similar or different**.
  - Contrastive learning can also be combined with labels, i.e. GE2E Loss.

- **How do we define similar datapoints?**
- **Standard approach: Create similar datapoints**
  - **Augmentation**.
  - **Split and duplicate**.
- **Our approach: Use structural information** about the dataset.
  - Data collection **time information**: utterances collected from an Alexa device within a short time period are mostly from a single speaker ➔ self-supervised speaker recognition

Photo by Raquel Martínez on Unsplash

Approach and Contributions

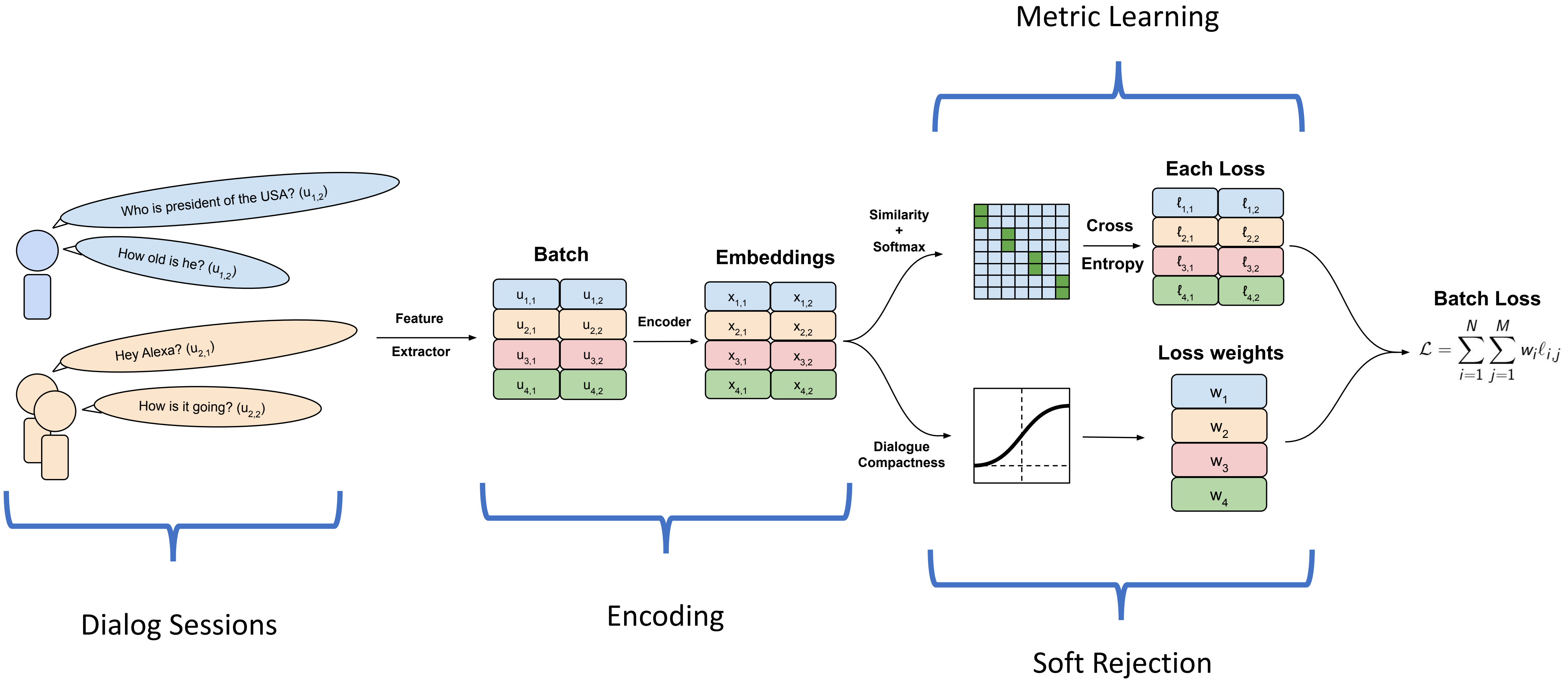
The Big Picture

Dialogue dataset from human-device interactions is an **alternative unlabeled data source** that can be leveraged for **speaker recognition model** pretraining.

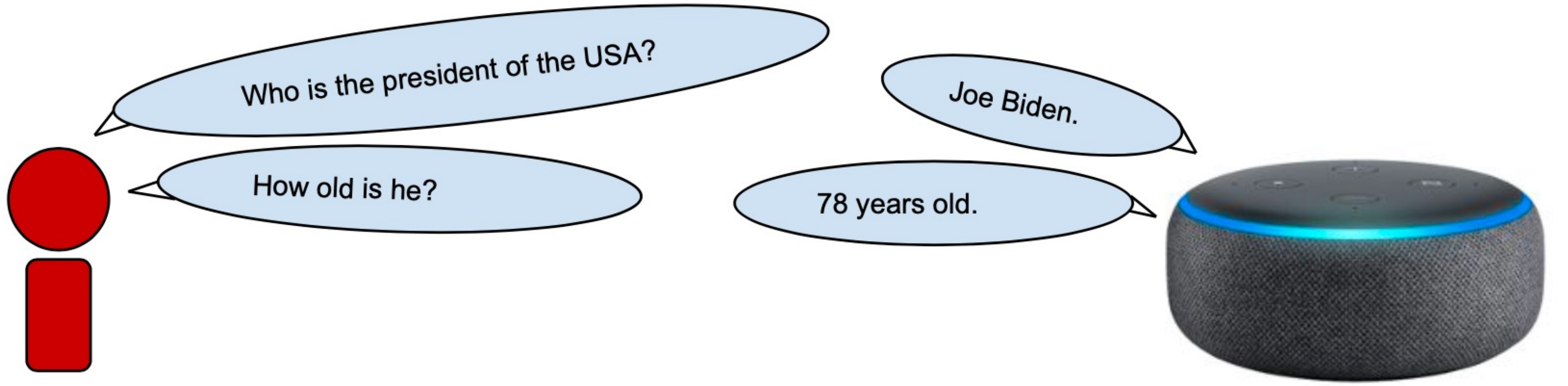
Key Technical Components

1. **Extracting positive and negative pairs from unlabeled Alexa dialogue sessions:** Utterances within a dialogue session provide positive pairs. Utterances from different devices provide negative pairs.
2. **Self-supervised soft rejection:** Dialogue “compactness” measure to reject incorrect/noisy positive pairs (e.g, arising due to multiple speakers).
3. **Fine-tuning:** fine-tuning the pretrained model on a **small labeled dataset** yields results comparable to fully-supervised training on a much larger dataset.

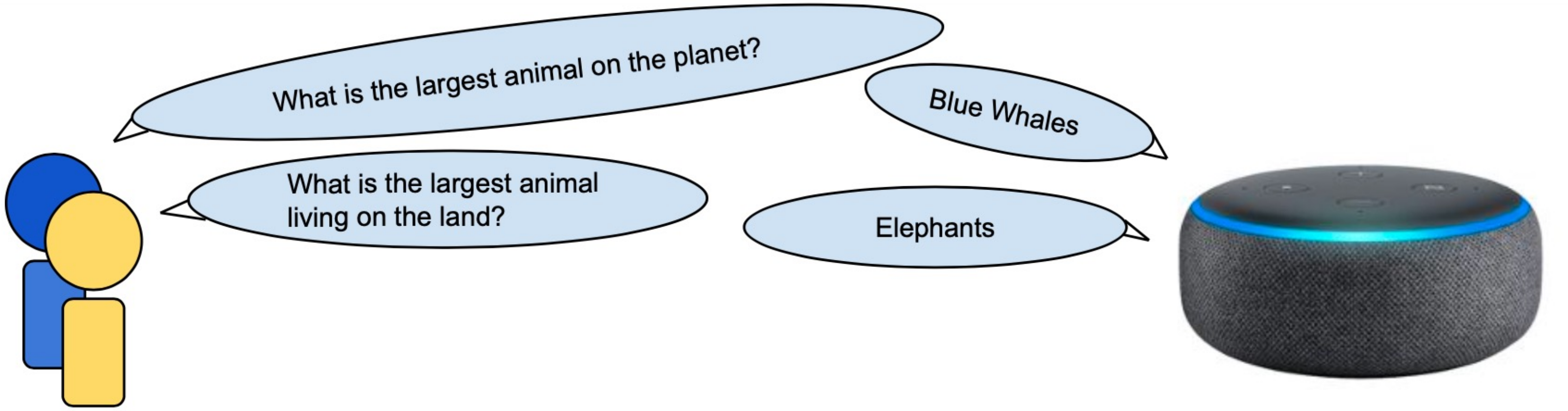
Proposed Framework



Alexa Dialogue Sessions



Single Speaker
(Most of the dialogues)



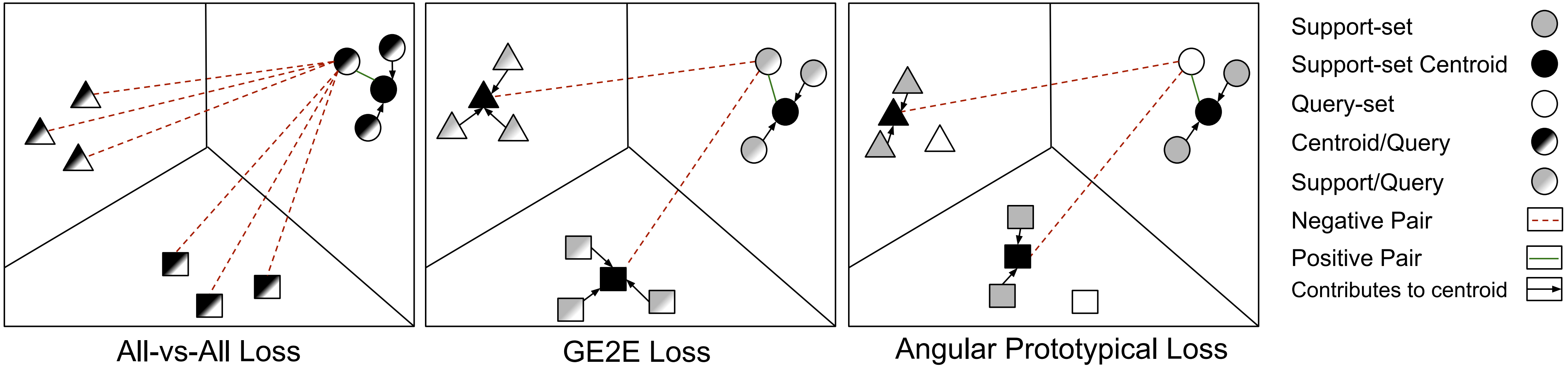
Multi Speaker
(Some of the dialogues)

Pretraining and Evaluation Dataset

- **Alexa Dialogue Dataset (Pretraining)**
  - **De-identified speech dialogues** from Alexa devices.
  - **927,000 dialogues -> 1800 hours of speech data**.
- **Annotated Alexa Dataset (Evaluation)**
  - Randomly sampled **de-identified utterances from a year's traffic**.
  - **Multiple human annotators**.
  - We only use samples with **consistent annotation**.
  - We report the **Equal Error Rate (EER) reduction** values for models.

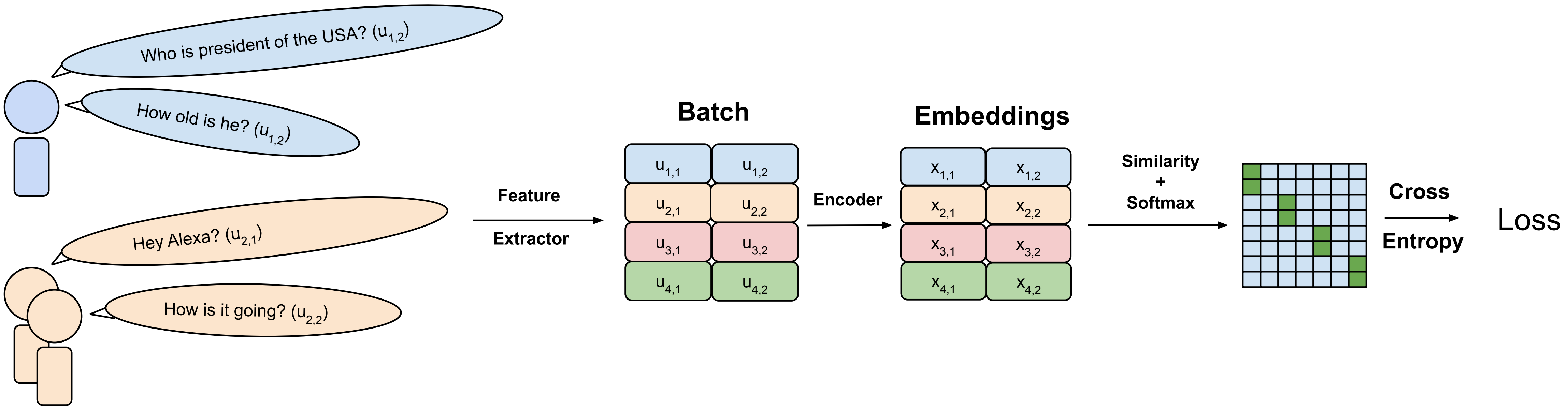
Loss Functions

What is the best loss function for our problem?



- **Some dialogues** may contain utterances from **different speakers**.
- **All-versus-all (AvA):**
  - avoiding the **flawed centroid problem**.
  - **increasing** the effective number of **negative pairs**.

Naive Framework for Self-Supervised Training



- We get two utterances from each dialogue.
- Compute embeddings using the encoder model.

Results of Pretraining

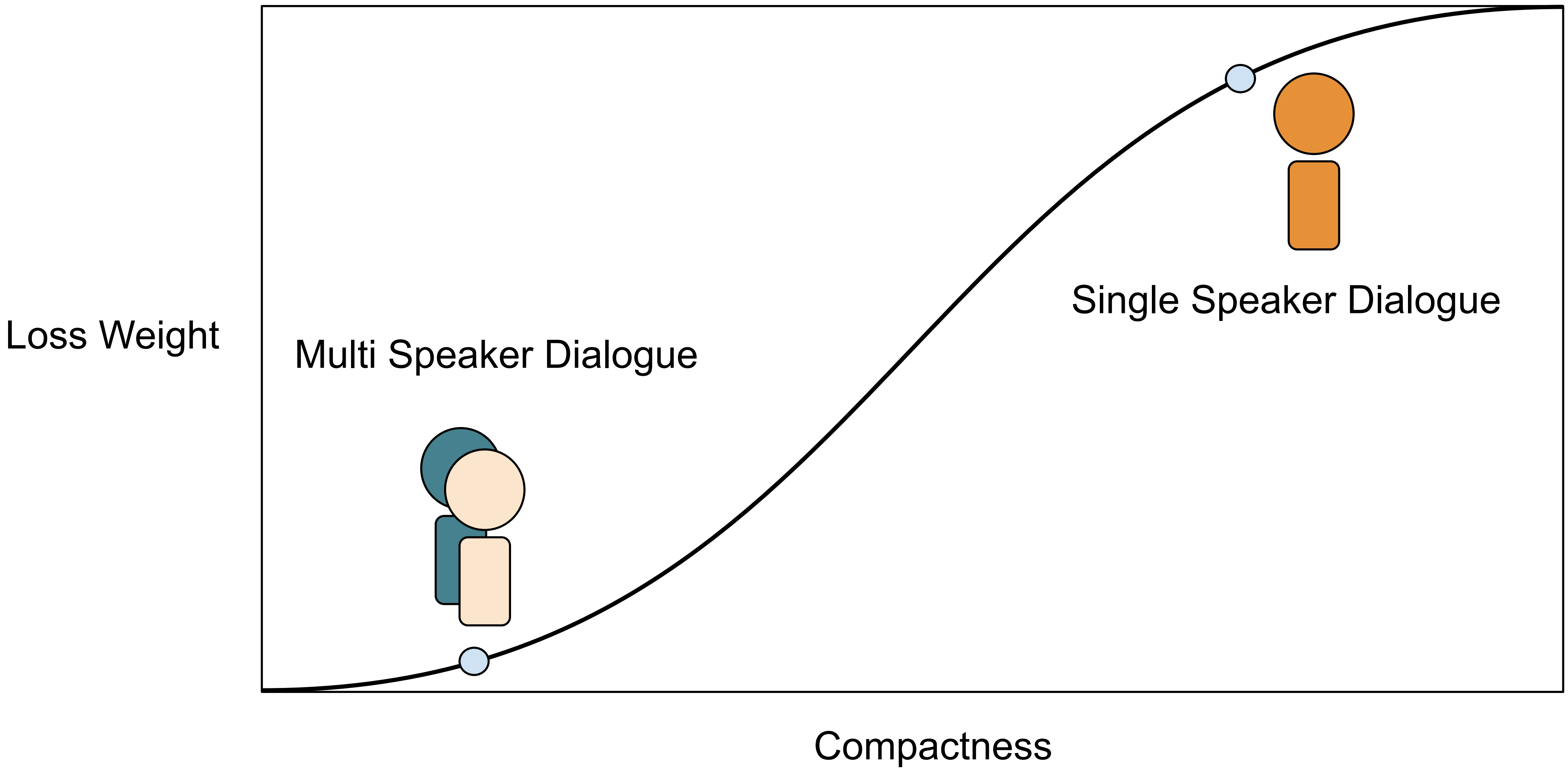
amazon alexa

Batch size = 256 → 256 Dialogues

Training Data	Method type	Loss Function	EER Reduction
VoxCeleb2	Supervised	GE2E	0.0%
Alexa-Dialogue	Self-supervised	AvA	**+19.32%**
Alexa-Dialogue	Self-supervised	GE2E	+18.36%
Alexa-Dialogue	Self-supervised	A-Proto	+18.78%

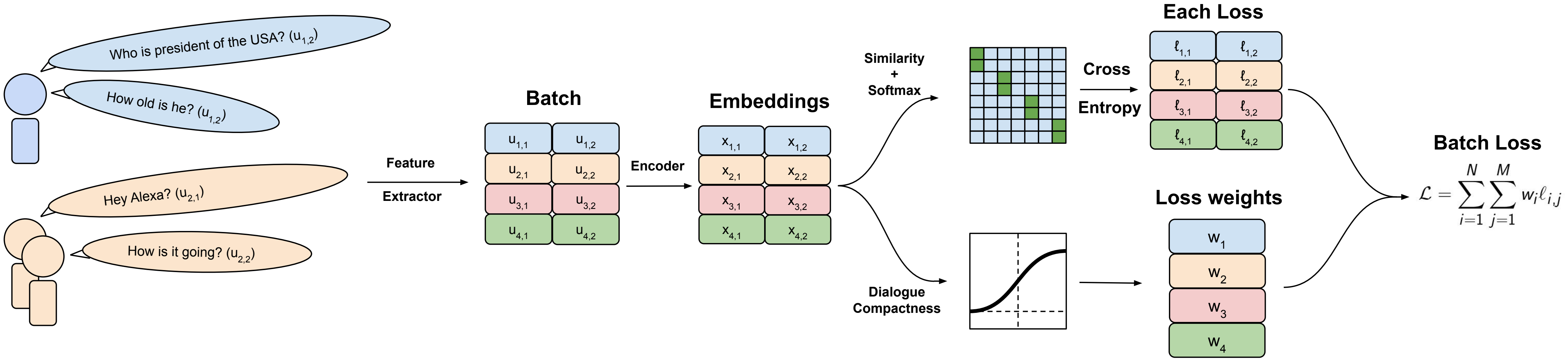
- Neural network **learns speaker ID related features**.
- Can we improve these results?
  - How to **reduce the impact** of **multi-speaker dialogues** in the learning?

Soft Rejection Mechanism



IDEA: Reduce the effect of multi-speaker dialogues on learning by **lowering the loss contribution** from the **dialogues with lower compactness scores**.

General Framework for Self-Supervised Training



- We incorporate the **soft rejection mechanism** to **eliminate multi-speaker dialogues** along the way, without supervision.

Results with Soft Rejection

Loss	Batch Size			
	32	64	128	256
All-vs-All	0.00%	+2.91%	+6.56%	**+8.20%**
Rejection + All-vs-All	**+3.76%**	+7.65%	+18.76%	**+19.00%**
A-Proto	0.00%	+7.32%	+8.52%	+12.93%
Rejection + A-proto	+7.55%	+12.58%	+16.76%	+25.85%
GE2E	0.00%	+3.24%	+3.06%	+6.36%
Rejection + GE2E	+10.64%	+17.75%	+17.99%	+13.83%

- **Soft Rejection** mechanism **Improves EER** consistently for all three loss functions for different batch sizes.
- Helping the model **focus on clean dialogues** rather than noisy ones.

Fine-Tuning Dataset

amazon alexa

We fine-tune the pretrained network on different **labeled Alexa datasets** with varying number of speakers, where **the total utterance duration for a speaker** is around **150 seconds** on average.

- 1024 different speakers → 150,000 seconds of utterances
- 2048 different speakers → 300,000 seconds of utterances
- 4096 different speakers → 600,000 seconds of utterances
- 8192 different speakers → 1,200,000 seconds of utterances

Results

amazon alexa

Pretraining	Loss	Episodes	Labeled Dataset Speaker Count			
			1,024	2,048	4,096	8,192
-	GE2E	1000	0.00%	0.00%	0.00%	0.00%

- Baseline: **model trained from scratch using GE2E loss** for 1000 episodes.

Results

Pretraining	Loss	Episodes	Labeled Dataset Speaker Count			
			1,024	2,048	4,096	8,192
-	GE2E	1000	0.00%	0.00%	0.00%	0.00%
COLA	GE2E	300	-8.81%	-23.57%	-37.07%	-44.21%
APC	GE2E	300	+24.34%	+23.13%	+19.48%	+15.35%
VoxCeleb2	GE2E	300	+31.38%	+25.91%	+20.95%	+15.61%

- **COLA[1]** framework **does not provide a good pretraining mechanism**.
- **Voxceleb2** and **APC[2]** frameworks **improve performance** with the learned representations.

[1] Saeed, Aaqib, David Grangier, and Neil Zeghidour. "Contrastive learning of general-purpose audio representations." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.
[2] Chung, Yu-An, and James Glass. "Generative pre-training for speech with autoregressive predictive coding." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

Results

amazon alexa

Pretraining	Loss	Episodes	Labeled Dataset Speaker Count			
			1,024	2,048	4,096	8,192
-	GE2E	1000	0.00%	0.00%	0.00%	0.00%
COLA	GE2E	300	-8.81%	-23.57%	-37.07%	-44.21%
APC	GE2E	300	+24.34%	+23.13%	+19.48%	+15.35%
VoxCeleb2	GE2E	300	+31.38%	+25.91%	+20.95%	+15.61%
Dialogue+AvA (ours)	GE2E	300	+40.18%	+34.19%	**+31.10%**	**+27.10%**
Dialogue+A-Proto (ours)	GE2E	300	**+41.28%**	**+34.77%**	+30.03%	+26.57%
Dialogue+GE2E (ours)	GE2E	300	+40.12%	+32.86%	+27.49%	+23.42%

- **Dialogue pretraining outperforms** all the other pretraining methods compared with.

Conclusions

- **Temporal proximity** provides a **valuable pseudo-label** which can be leveraged to learn speaker-ID related features.
- A **self-supervised soft rejection** mechanism is **very effective** to deal with false positive pair problem in this context.

Future Work: Exploring the interaction between labels and self-supervision

- Is self-supervised pretraining still useful if we have access to a large labeled dataset?
- Can adding a small labeled dataset to self-supervised pretraining improve focus on speaker ID?