SemanticVAD

Fengshi Teng

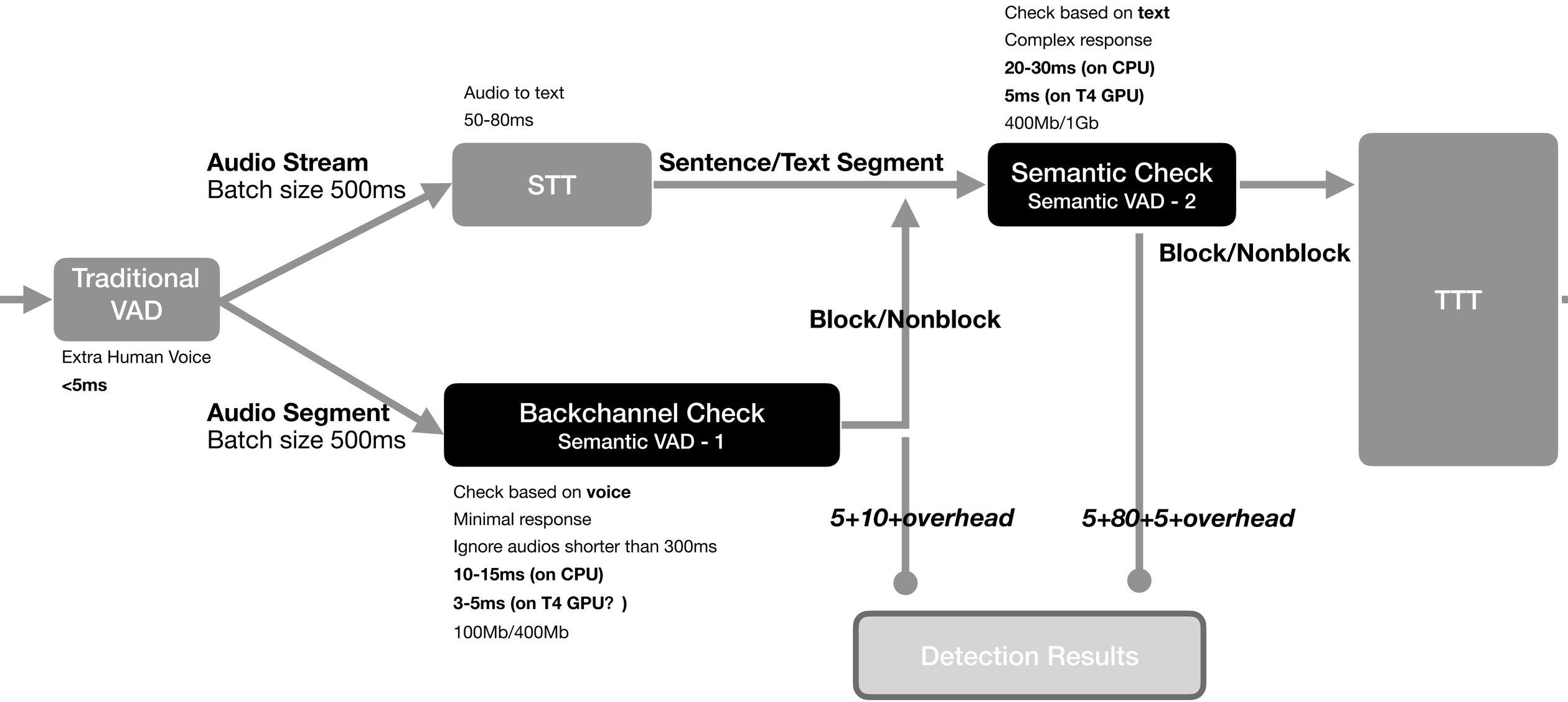
Task & Goal
Schema & Approach
Finetuned Models
Performance
Deployment & Cost

TASK

Fine-tune a model to construct a **semantic voice activity detector (VAD)** capable of identifying backchannels or interruptions, so as to reduce unnecessary disruptions in a spoken dialogue system.

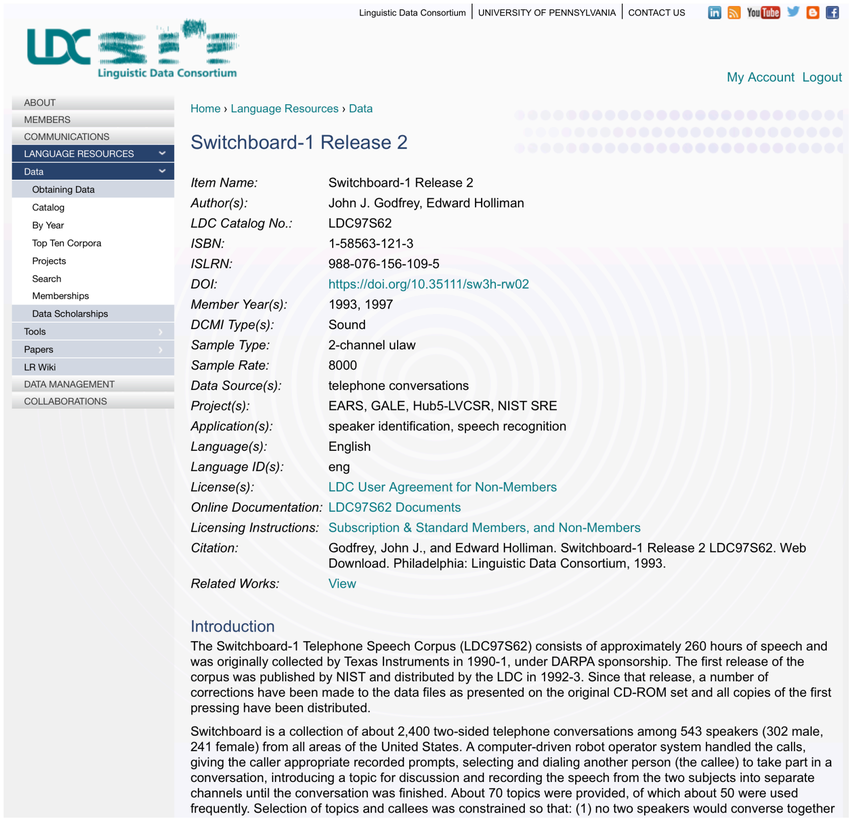
Goal: detect interruptions within **200ms** while maintaining efficient memory usage and limited budget constraints.

SCHEMA


Check based on **text**
Complex response
20-30ms (on CPU)
5ms (on T4 GPU)
400Mb/1Gb
Audio to text
50-80ms
Audio Stream
Batch size 500ms
STT
Sentence/Text Segment
Semantic Check
Semantic VAD - 2
Block/Nonblock
TTT
Traditional VAD
Extra Human Voice
<5ms
Block/Nonblock
Audio Segment
Batch size 500ms
Backchannel Check
Semantic VAD - 1
Check based on **voice**
Minimal response
Ignore audios shorter than 300ms
10-15ms (on CPU)
3-5ms (on T4 GPU?)
100Mb/400Mb
5+10+overhead
5+80+5+overhead
Detection Results

DATASETS

Train/Test Dataset



Linguistic Data Consortium | UNIVERSITY OF PENNSYLVANIA | CONTACT US

Linguistic Data Consortium

My Account Logout

ABOUT
MEMBERS
COMMUNICATIONS
LANGUAGE RESOURCES
Data
Obtaining Data
Catalog
By Year
Top Ten Corpora
Projects
Search
Memberships
Data Scholarships
Tools
Papers
LR Wiki
DATA MANAGEMENT
COLLABORATIONS

Home › Language Resources › Data

Switchboard-1 Release 2

Item Name:	Switchboard-1 Release 2
Author(s):	John J. Godfrey, Edward Holliman
LDC Catalog No.:	LDC97S62
ISBN:	1-58563-121-3
ISLRN:	988-076-156-109-5
DOI:	https://doi.org/10.35111/sw3h-rw02
Member Year(s):	1993, 1997
DCMI Type(s):	Sound
Sample Type:	2-channel ulaw
Sample Rate:	8000
Data Source(s):	telephone conversations
Project(s):	EARS, GALE, Hub5-LVCSR, NIST SRE
Application(s):	speaker identification, speech recognition
Language(s):	English
Language ID(s):	eng
License(s):	LDC User Agreement for Non-Members
Online Documentation:	LDC97S62 Documents
Licensing Instructions:	Subscription & Standard Members, and Non-Members
Citation:	Godfrey, John J., and Edward Holliman. Switchboard-1 Release 2 LDC97S62. Web Download. Philadelphia: Linguistic Data Consortium, 1993.
Related Works:	View

Introduction

The Switchboard-1 Telephone Speech Corpus (LDC97S62) consists of approximately 260 hours of speech and was originally collected by Texas Instruments in 1990-1, under DARPA sponsorship. The first release of the corpus was published by NIST and distributed by the LDC in 1992-3. Since that release, a number of corrections have been made to the data files as presented on the original CD-ROM set and all copies of the first pressing have been distributed.

Switchboard is a collection of about 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from all areas of the United States. A computer-driven robot operator system handled the calls, giving the caller appropriate recorded prompts, selecting and dialing another person (the callee) to take part in a conversation, introducing a topic for discussion and recording the speech from the two subjects into separate channels until the conversation was finished. About 70 topics were provided, of which about 50 were used frequently. Selection of topics and callees was constrained so that: (1) no two speakers would converse together

Backchannels

500+
2-channel telephone recordings
sph files labelled
20GB

```
sw3704A-ms98-a-0001 0.000000 0.447125 [noise]
sw3704A-ms98-a-0002 0.447125 2.618875 what was the topic [laughter]
sw3704A-ms98-a-0003 2.618875 4.062125 [silence]
sw3704A-ms98-a-0004 4.062125 5.737500 i said uh
sw3704A-ms98-a-0005 5.737500 7.461625 r[ight]- right
sw3704A-ms98-a-0006 7.461625 8.556250 [silence]
sw3704A-ms98-a-0007 8.556250 10.088625 yeah
sw3704A-ms98-a-0008 10.088625 16.820375 [silence]
sw3704A-ms98-a-0009 16.820375 20.142500 well that's a good one i hadn't thought about that yeah
sw3704A-ms98-a-0010 20.142500 23.932625 [silence]
sw3704A-ms98-a-0011 23.932625 25.436125 yeah really
sw3704A-ms98-a-0012 25.436125 31.726250 [silence]
sw3704A-ms98-a-0013 31.726250 33.595375 uh oh
sw3704A-ms98-a-0014 33.595375 36.819750 there there was a stigma attached to it
sw3704A-ms98-a-0015 36.819750 40.785750 [silence]
sw3704A-ms98-a-0016 40.785750 43.452500 [vocalized-noise] an[d]- and a necessity
sw3704A-ms98-a-0017 43.452500 44.947625 yeah
sw3704A-ms98-a-0018 44.947625 58.169375 [silence]
.....
```

Extract

```
"uh-huh", "mm-hm", "mhm", "yeah", "yeahhh", "right", "okay", "oh", "yes",
    "uh", "um", "huh", "hm", "i", "alright", "sure", "really", "wow", "hmm", "ah", "eh",
```

Head of long speech
(500ms)

Label 1 40000+

Backchannels
(500ms)

Label 0 30000+

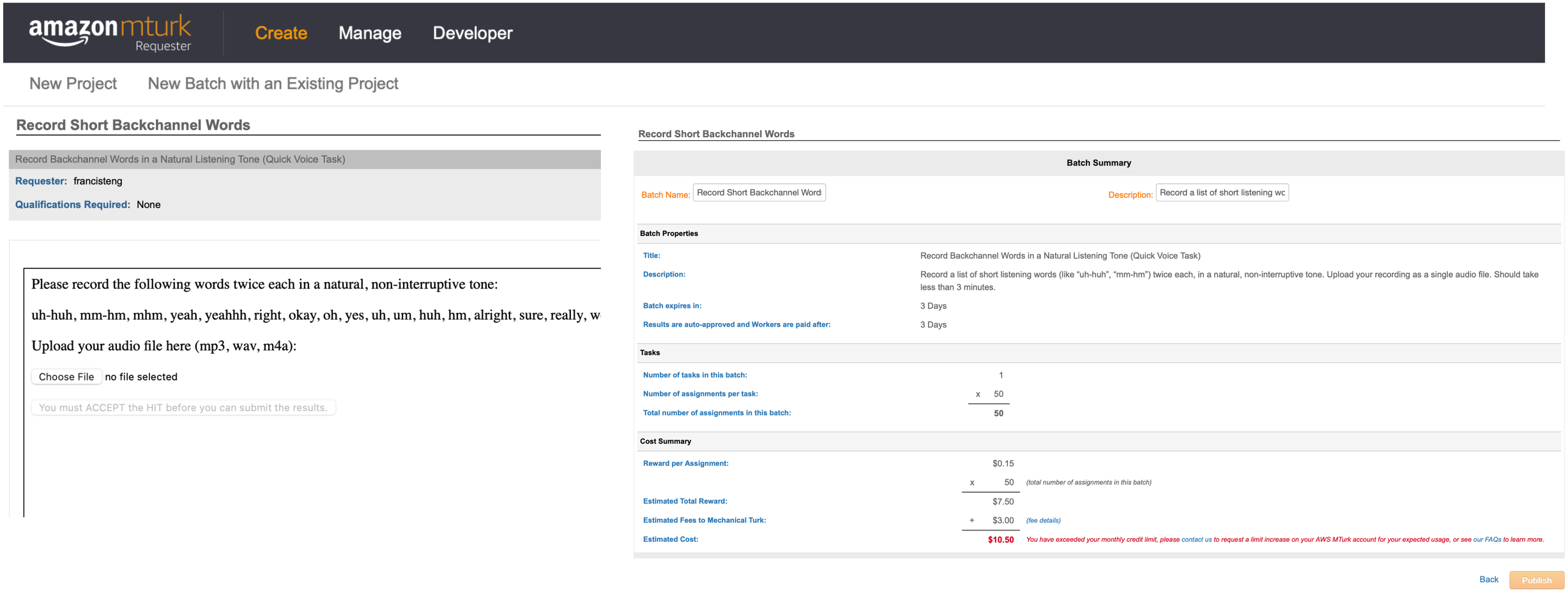
Mix&Shuffle

Training Set (train+validation) + Test Set

30000+ 20000+

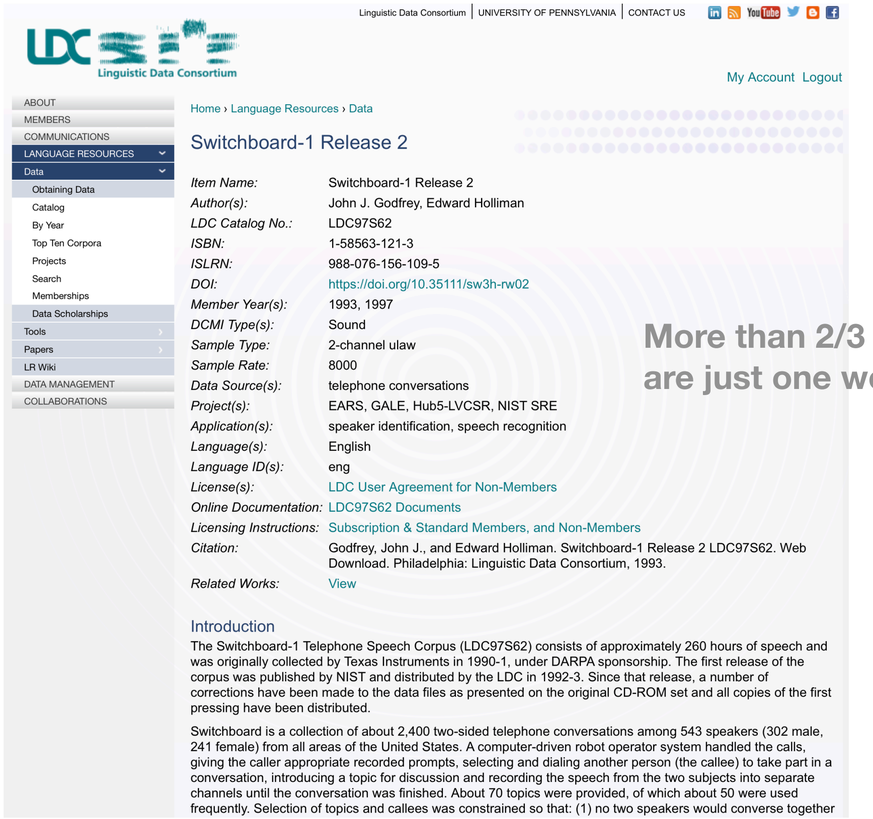
DATASETS

Generalization Test Dataset



DATASETS

Train/Test Dataset



Linguistic Data Consortium | UNIVERSITY OF PENNSYLVANIA | CONTACT US

Linguistic Data Consortium

My Account Logout

ABOUT
MEMBERS
COMMUNICATIONS
LANGUAGE RESOURCES
Data
Obtaining Data
Catalog
By Year
Top Ten Corpora
Projects
Search
Memberships
Data Scholarships
Tools
Papers
LR Wiki
DATA MANAGEMENT
COLLABORATIONS

Home › Language Resources › Data

Switchboard-1 Release 2

Item Name:	Switchboard-1 Release 2
Author(s):	John J. Godfrey, Edward Holliman
LDC Catalog No.:	LDC97S62
ISBN:	1-58563-121-3
ISLRN:	988-076-156-109-5
DOI:	https://doi.org/10.35111/sw3h-rw02
Member Year(s):	1993, 1997
DCMI Type(s):	Sound
Sample Type:	2-channel ulaw
Sample Rate:	8000
Data Source(s):	telephone conversations
Project(s):	EARS, GALE, Hub5-LVCSR, NIST SRE
Application(s):	speaker identification, speech recognition
Language(s):	English
Language ID(s):	eng
License(s):	LDC User Agreement for Non-Members
Online Documentation:	LDC97S62 Documents
Licensing Instructions:	Subscription & Standard Members, and Non-Members
Citation:	Godfrey, John J., and Edward Holliman. Switchboard-1 Release 2 LDC97S62. Web Download. Philadelphia: Linguistic Data Consortium, 1993.
Related Works:	View

Introduction

The Switchboard-1 Telephone Speech Corpus (LDC97S62) consists of approximately 260 hours of speech and was originally collected by Texas Instruments in 1990-1, under DARPA sponsorship. The first release of the corpus was published by NIST and distributed by the LDC in 1992-3. Since that release, a number of corrections have been made to the data files as presented on the original CD-ROM set and all copies of the first pressing have been distributed.

Switchboard is a collection of about 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from all areas of the United States. A computer-driven robot operator system handled the calls, giving the caller appropriate recorded prompts, selecting and dialing another person (the callee) to take part in a conversation, introducing a topic for discussion and recording the speech from the two subjects into separate channels until the conversation was finished. About 70 topics were provided, of which about 50 were used frequently. Selection of topics and callees was constrained so that: (1) no two speakers would converse together

Backchannels
500+
2-channel telephone recordings
sph files labelled
20GB

```
sw3704A-ms98-a-0001 0.000000 0.447125 [noise]
sw3704A-ms98-a-0002 0.447125 2.618875 what was the topic [laughter]
sw3704A-ms98-a-0003 2.618875 4.062125 [silence]
sw3704A-ms98-a-0004 4.062125 5.737500 i said uh
sw3704A-ms98-a-0005 5.737500 7.461625 r[ight]- right
sw3704A-ms98-a-0006 7.461625 8.556250 [silence]
sw3704A-ms98-a-0007 8.556250 10.088625 yeah
sw3704A-ms98-a-0008 10.088625 16.820375 [silence]
sw3704A-ms98-a-0009 16.820375 20.142500 well that's a good one i hadn't thought about that yeah
sw3704A-ms98-a-0010 20.142500 23.932625 [silence]
sw3704A-ms98-a-0011 23.932625 25.436125 yeah really
sw3704A-ms98-a-0012 25.436125 31.726250 [silence]
sw3704A-ms98-a-0013 31.726250 33.595375 uh oh
sw3704A-ms98-a-0014 33.595375 36.819750 there there was a stigma attached to it
sw3704A-ms98-a-0015 36.819750 40.785750 [silence]
sw3704A-ms98-a-0016 40.785750 43.452500 [vocalized-noise] an[d]- and a necessity
sw3704A-ms98-a-0017 43.452500 44.947625 yeah
sw3704A-ms98-a-0018 44.947625 58.169375 [silence]
.....
```

Extract / Filter

More than 2/3 of the backchannels are just one word/ monosyllabic word!

More than 1/3 of a speech segment starts with backchannels!

2587 4189

Backchannels (2-10 words) **Label 1** 700+

\+

Interruption (2-15 words) **Label 0** 700+

Mix&Shuffle

Generated by AI

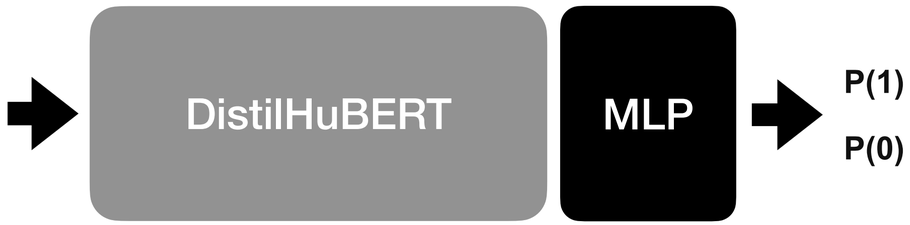
Training Set (train+validation)
1000/5000

Test Set
1500

MODELS: SemanticVAD1

Base model

- **Model Name**: DistilHuBERT
- **Developed by**: NTU-SPML
- **Architecture**: 6-layer Transformer encoder (compared to 12 layers in HuBERT Base)
- **Input Format**: 16 kHz mono-channel audio
- **Pretraining Data**: Libri-Light 60k hours
- **Model Size**: Approximately 44 million parameters
- **Disk Size**: ~90 MB (400MB GPU Memory during runtime)
- **Inference Time (GPU)**: Approximately 9–12 milliseconds per second of audio



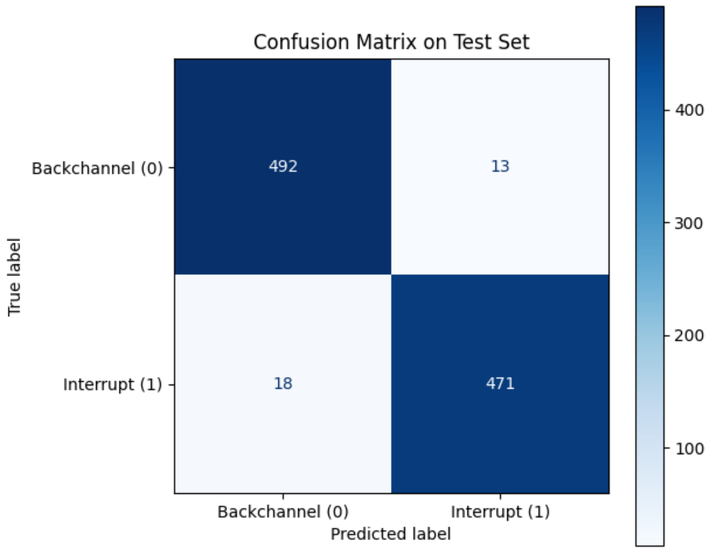
Finetune Constructor (MLP)

```python
from transformers.modeling_outputs import SequenceClassifierOutput

class DistilHuBERTClassifier(nn.Module):
    def __init__(self, base_model, num_labels):
        super().__init__()
        self.encoder = base_model
        # Use MLP
        self.classifier = nn.Sequential(
            nn.Linear(base_model.config.hidden_size, 512),
            nn.ReLU(),
            nn.Dropout(0.2),
            nn.Linear(512, 256),
            nn.ReLU(),
            nn.Dropout(0.2),
            nn.Linear(256, num_labels)
        )

    def forward(self, input_values, attention_mask=None, labels=None):
        outputs = self.encoder(input_values=input_values, attention_mask=attention_mask)
        pooled = outputs.last_hidden_state.mean(dim=1)
        logits = self.classifier(pooled)
        loss = None
        if labels is not None:
            loss = nn.CrossEntropyLoss()(logits, labels)
        return SequenceClassifierOutput(
            loss=loss,
            logits=logits
        )
```

MODELS: SemanticVAD1

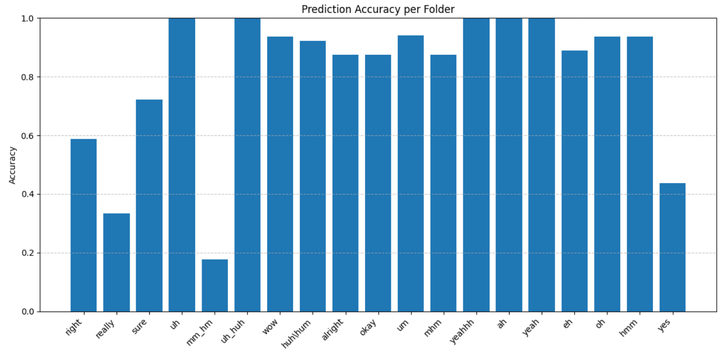
512*256*2

Accuracy on Test Dataset



✅ Overall Test Accuracy: 96.20%
🎯 Class 0 Accuracy: 96.85%
🎯 Class 1 Accuracy: 95.54%

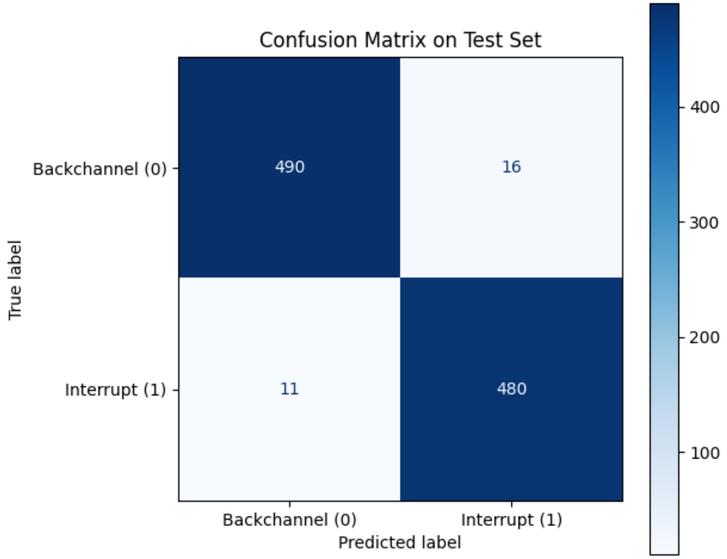
Accuracy on Generalization



✅ Overall Test Accuracy: 81.23%

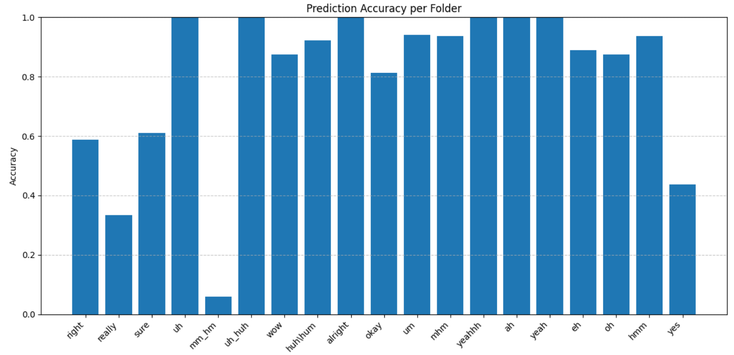
256*128*2

Accuracy on Test Dataset



✅ Overall Test Accuracy: 96.50%
🎯 Class 0 Accuracy: 97.44%
🎯 Class 1 Accuracy: 95.54%

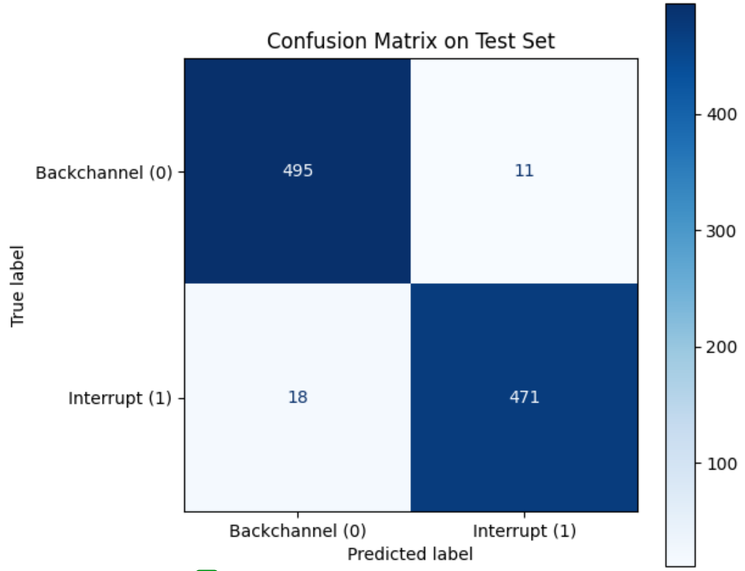
Accuracy on Generalization



✅ Overall Test Accuracy: 79.94%

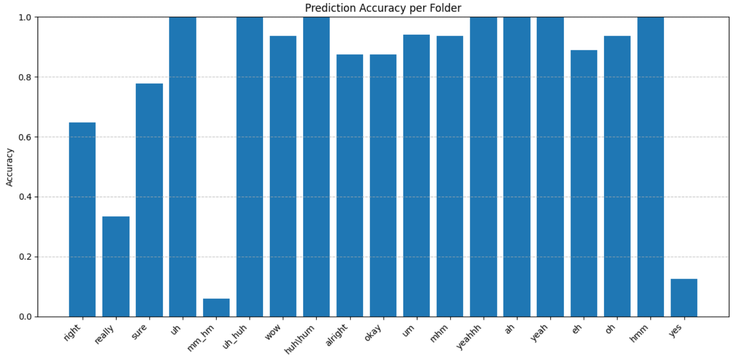
128*64*2

Accuracy on Test Dataset



✅ Overall Test Accuracy: 96.90%
🎯 Class 0 Accuracy: 96.46%
🎯 Class 1 Accuracy: 97.36%

Accuracy on Generalization

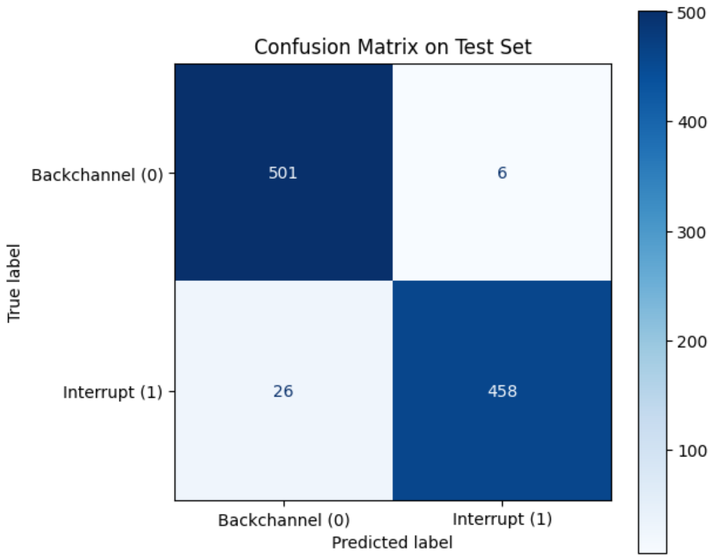


✅ Overall Test Accuracy: 80.58%

MODELS: SemanticVAD1

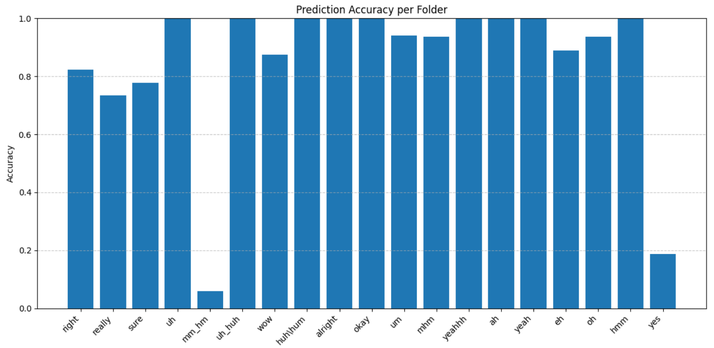
64*2

Accuracy on Test Dataset



✅ Overall Test Accuracy: 95.80%
🎯 Class 0 Accuracy: 98.62%
🎯 Class 1 Accuracy: 92.90%

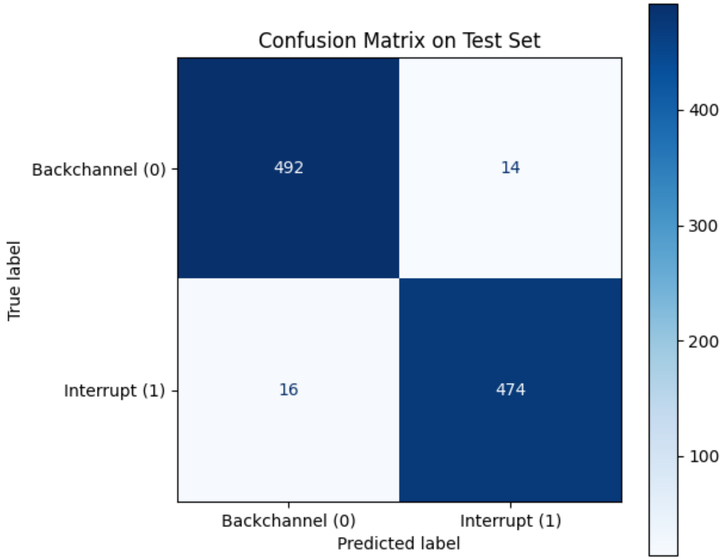
Accuracy on Generalization



✅ Overall Test Accuracy: 84.79%

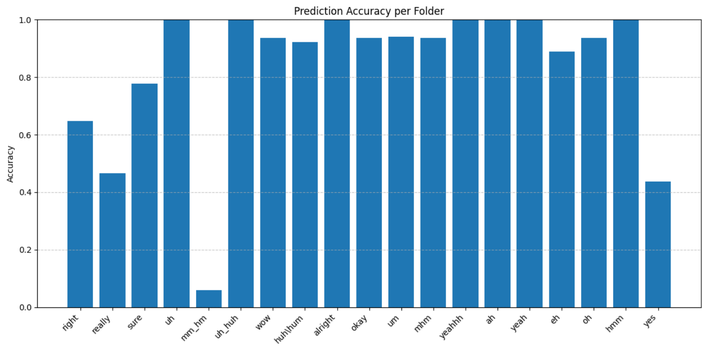
128*2

Accuracy on Test Dataset



✅ Overall Test Accuracy: 96.50%
🎯 Class 0 Accuracy: 96.86%
🎯 Class 1 Accuracy: 96.15%

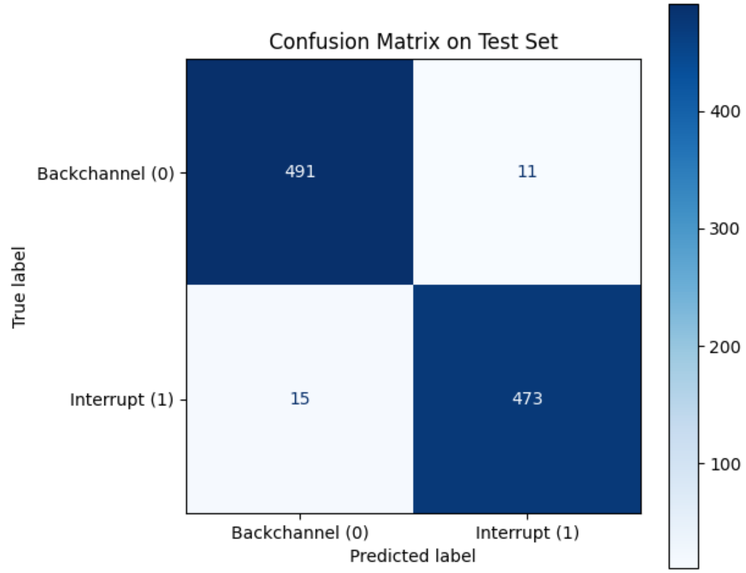
Accuracy on Generalization



✅ Overall Test Accuracy: 83.50%

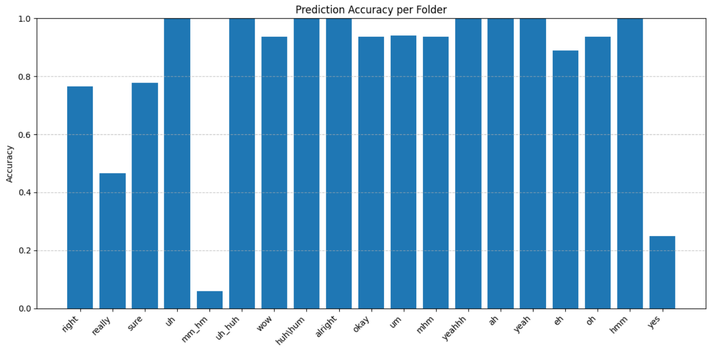
256*2

Accuracy on Test Dataset



✅ Overall Test Accuracy: 96.30%
🎯 Class 0 Accuracy: 96.65%
🎯 Class 1 Accuracy: 95.94%

Accuracy on Generalization

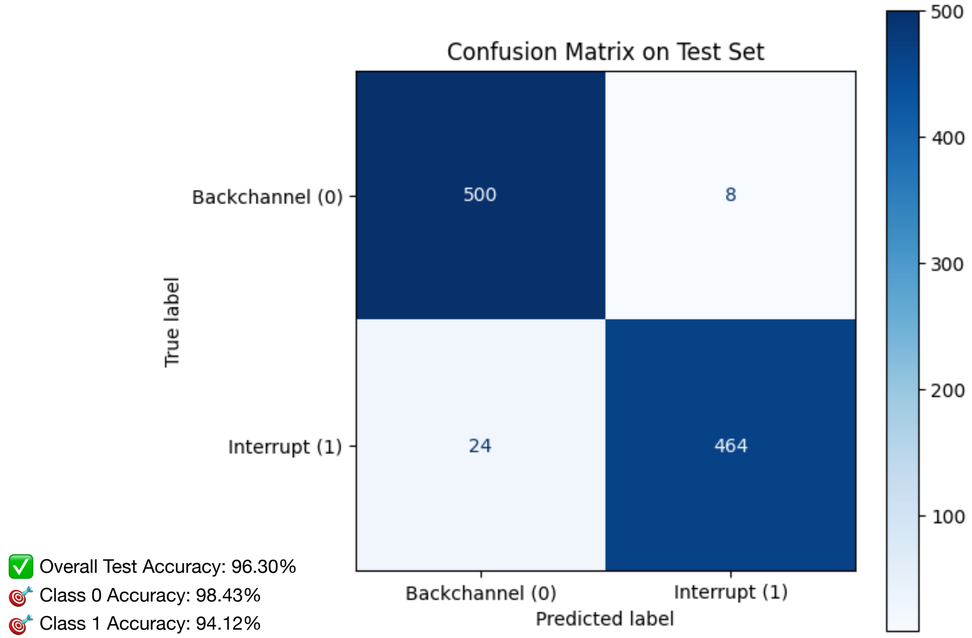


✅ Overall Test Accuracy: 83.50%

MODELS: SemanticVAD1

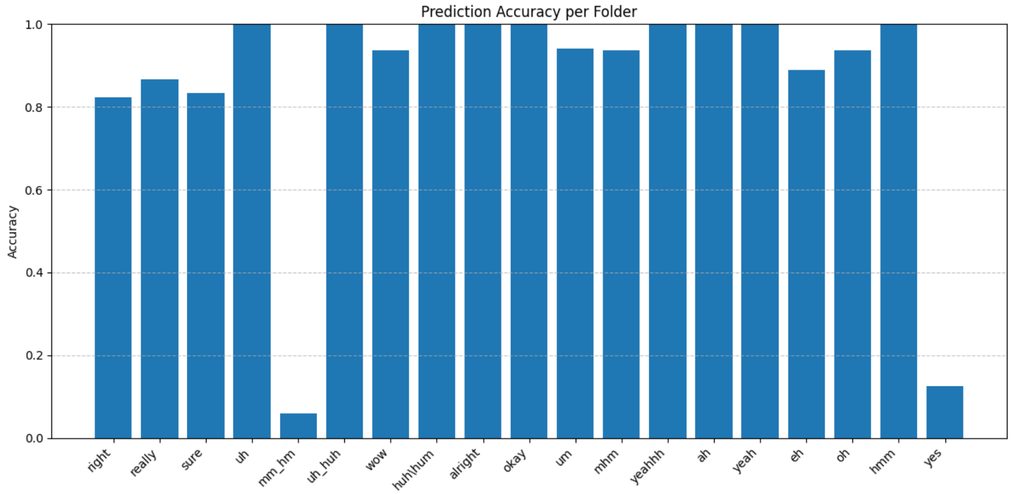
Just one linear layer is enough!

Accuracy on Test Dataset



✅ Overall Test Accuracy: 96.30%
🎯 Class 0 Accuracy: 98.43%
🎯 Class 1 Accuracy: 94.12%

Accuracy on Generalization



✅ Overall Test Accuracy: 85.76%

Performance varies on different words.

MODELS: SemanticVAD2

Base model

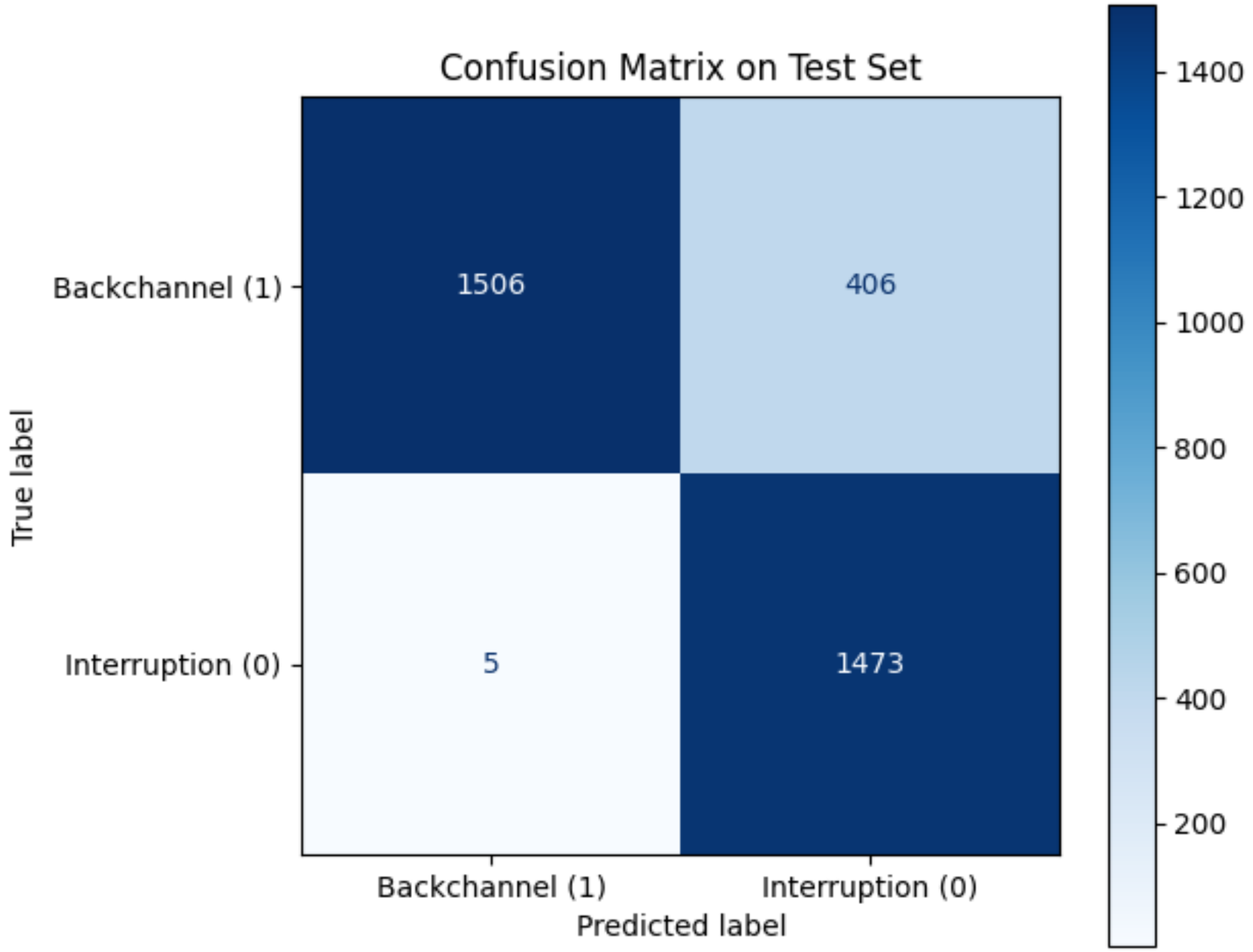
- **Model Name**: DistilBERT (distilbert-base-uncased)
- **Developed by**: Hugging Face
- **Architecture**: 6-layer Transformer encoder (compared to 12 layers in BERT Base)
- **Input Format**: Tokenized text (WordPiece, lowercased), max length 512
- **Pretraining Data**: English Wikipedia + BookCorpus (same as BERT)
- **Model Size**: Approximately 66 million parameters
- **Disk Size**: ~255 MB
- **Inference Time (GPU)**: Approximately 5-8 milliseconds per sentence (may vary depending on hardware)

Finetune Constructor (MLP)

```python
from torch import nn

class DistilBERTBackchannelScorer(nn.Module):
    def __init__(self, hidden_dim=768):
        super().__init__()
        self.encoder = DistilBertModel.from_pretrained("distilbert-base-uncased")
        self.classifier = nn.Sequential(
            nn.Linear(hidden_dim, 128),
            nn.ReLU(),
            nn.Dropout(0.2),
            nn.Linear(128, 1)
        )

    def forward(self, input_ids, attention_mask=None, labels=None):
        outputs = self.encoder(input_ids=input_ids, attention_mask=attention_mask)
        pooled = outputs.last_hidden_state[:, 0]  # CLS token
        logits = self.classifier(pooled).squeeze(-1)  # shape: (batch_size,)
        loss = None
        if labels is not None:
            labels = labels.float()  # BCE loss 要求 float
            loss = nn.BCEWithLogitsLoss()(logits, labels)
        return SequenceClassifierOutput(
            loss=loss,
            logits=logits
        )
```

MODELS: SemanticVAD2

Accuracy on Test Dataset

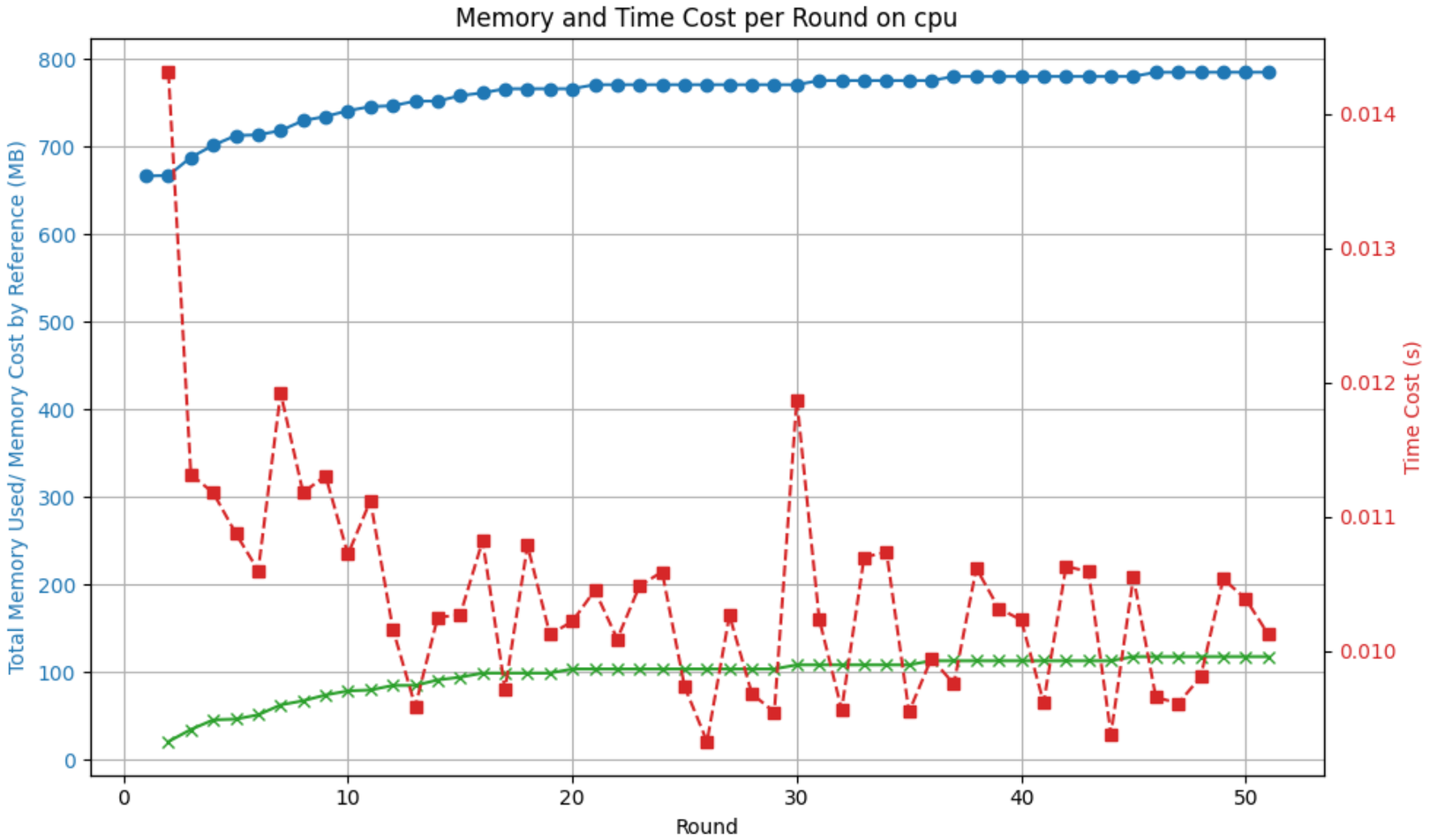


Model: basemodel*256*1

✅ Overall Test Accuracy: 87.88%
✅ F1 Score: 87.76%
🎯 Class 0 Accuracy: 78.77%
🎯 Class 1 Accuracy: 99.66%

DEPLOYMENT/COST

Memory USAGE
Semantic VAD 1
(on Mac, without CUDA)

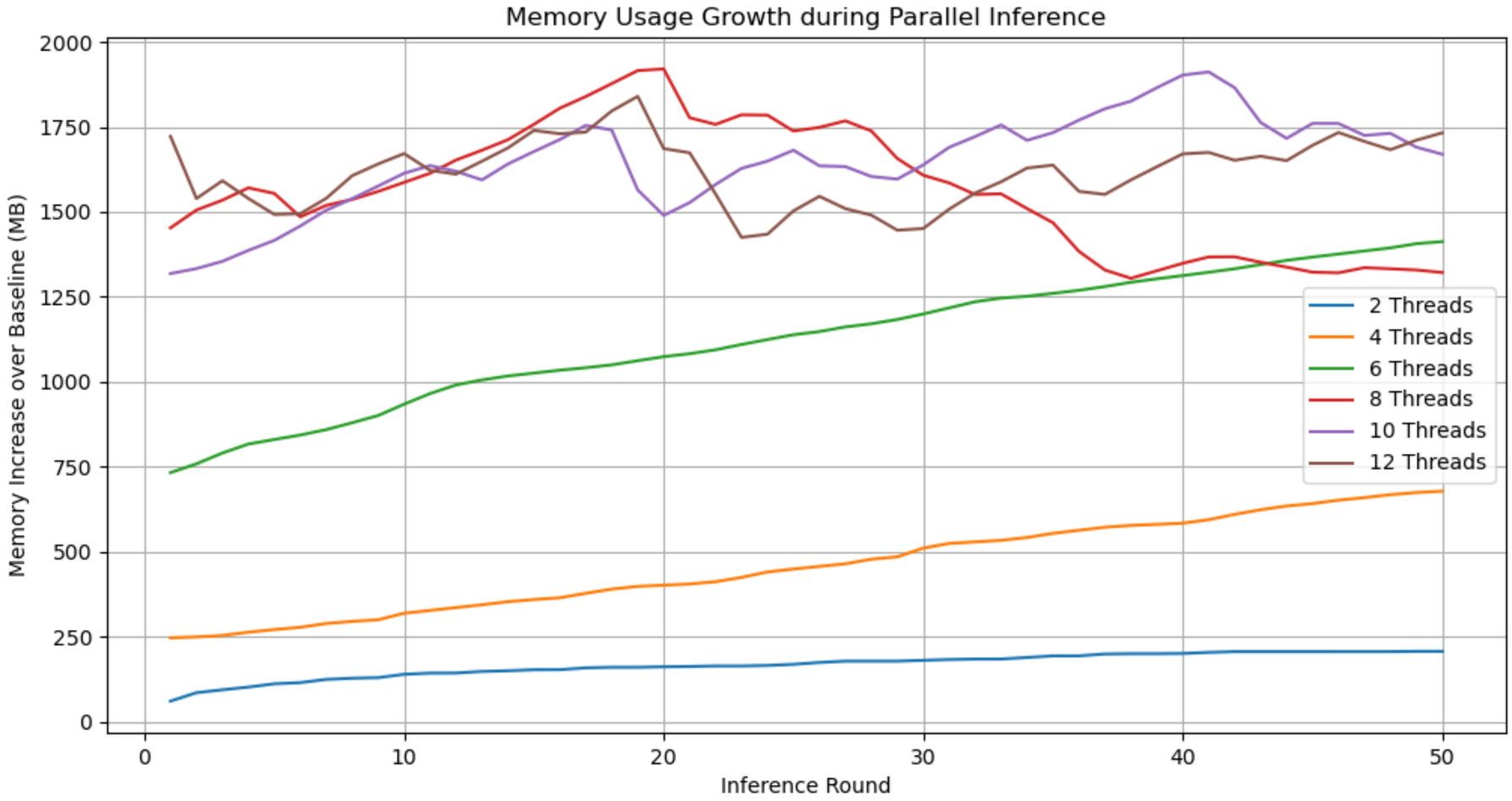


Total memory cost by reference: **117.62 MB**
Average input length: 0.500 s
Average time cost: **10.42 ms**

DEPLOYMENT/COST

Memory USAGE (Multithread) Semantic VAD 1

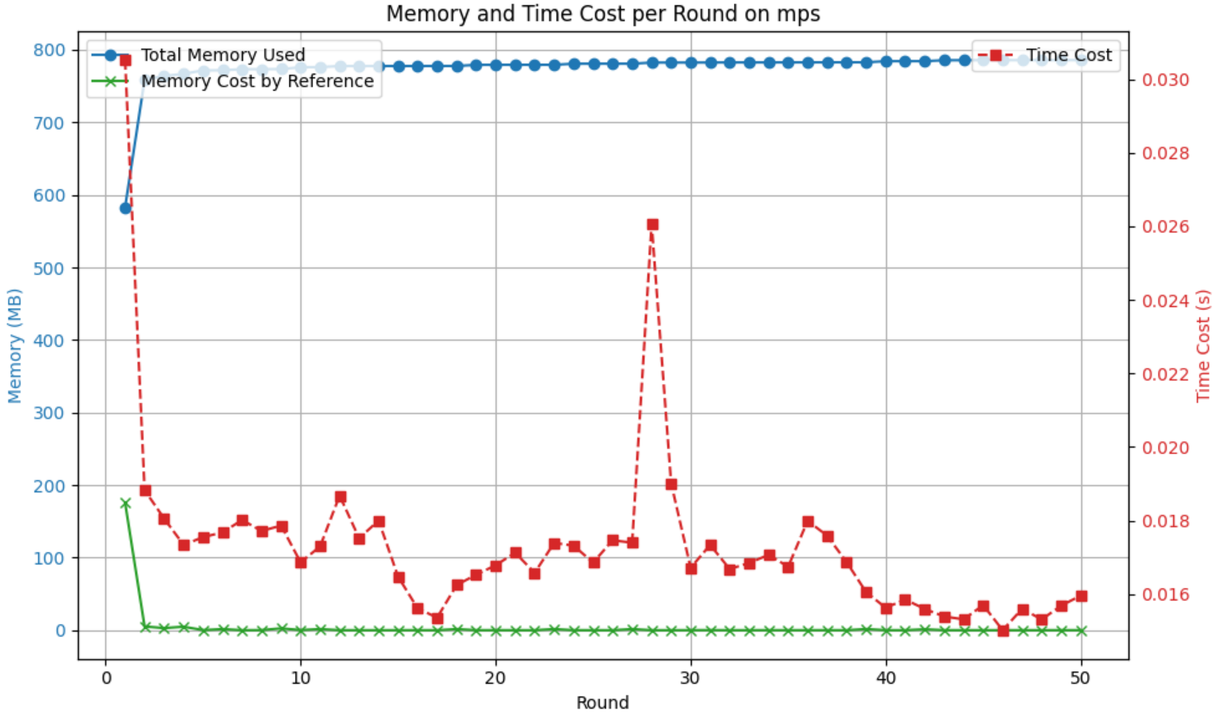
(on Mac, without CUDA)



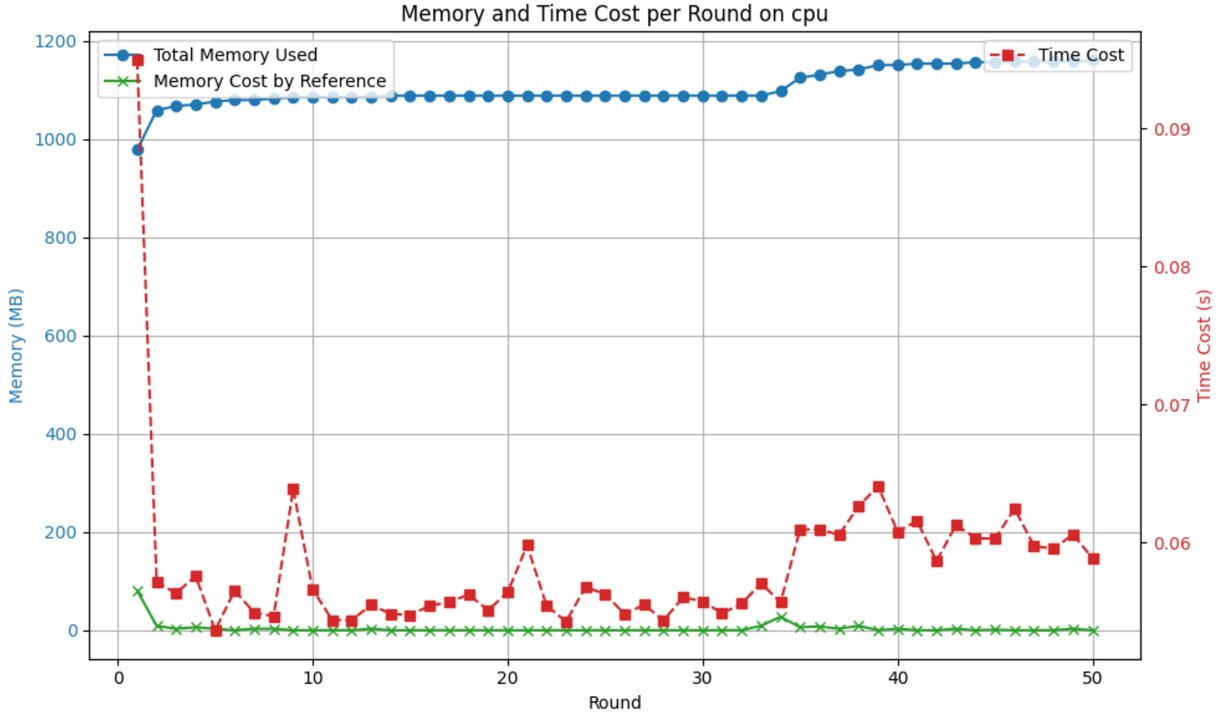
DEPLOYMENT/COST

Memory USAGE
Semantic VAD 2
(on Mac, without CUDA)

Total memory cost by reference: **192.78MB**
Average input length: <15 words
Average time cost:
Batch size 1: 17.30 ms
Batch size 8: 58.27ms



Batch size = 1

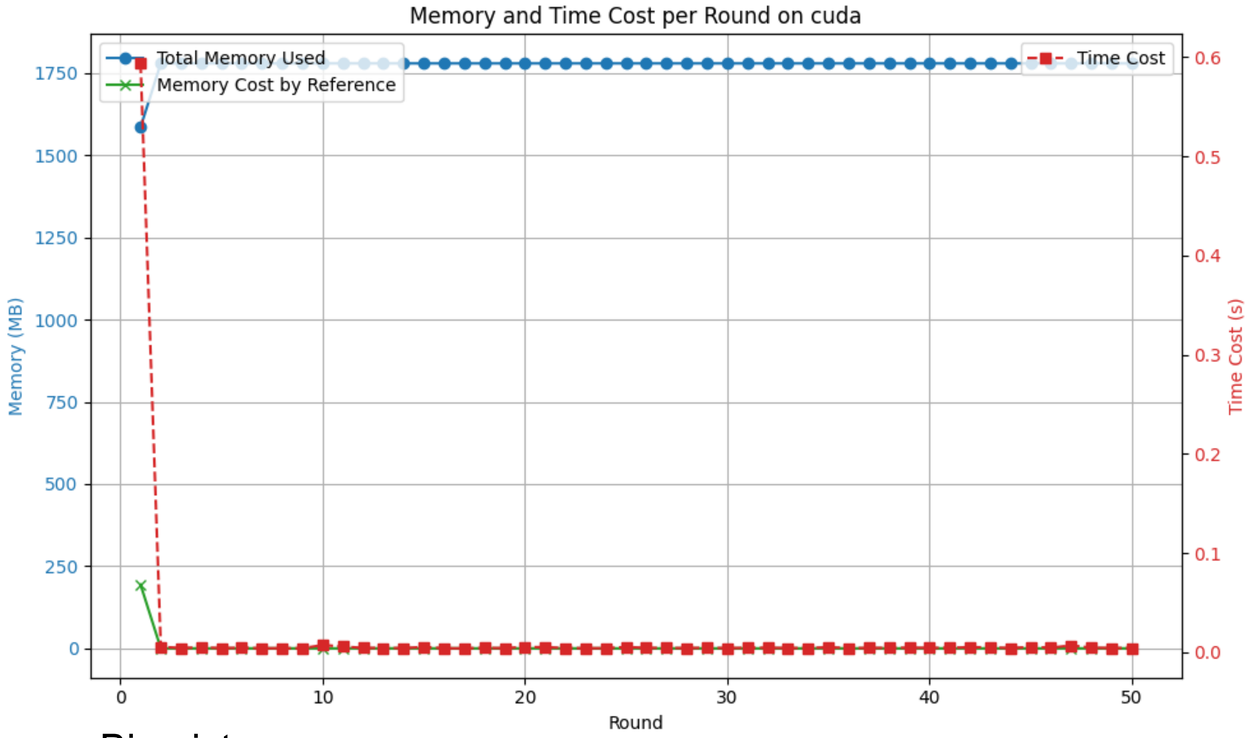


Batch size = 8

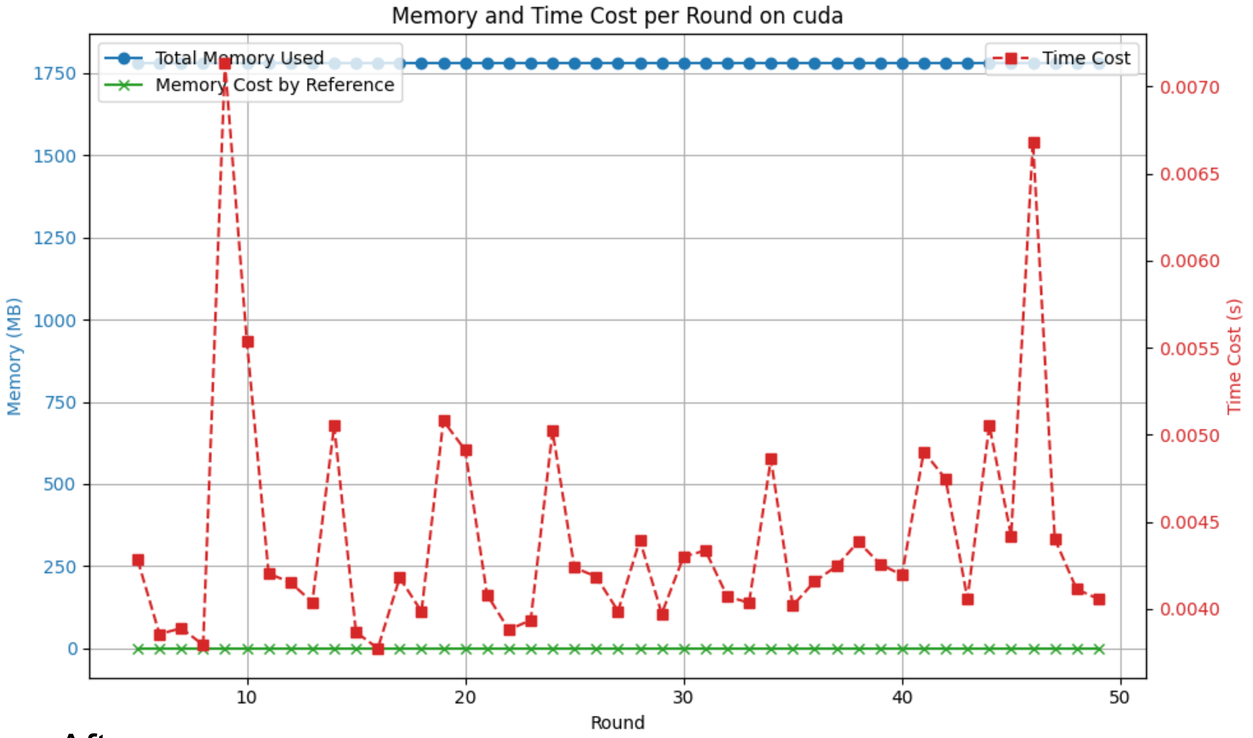
DEPLOYMENT/COST

Memory USAGE
Semantic VAD 2
(on Colab, with T4 CUDA)

Total memory cost by reference: **193.59MB**
Average input length: <15 words
Average time cost: (Batch size 1)
16.21ms (total average)
4.42ms (after warmup)



Big picture



After warmup

DEPLOYMENT/COST

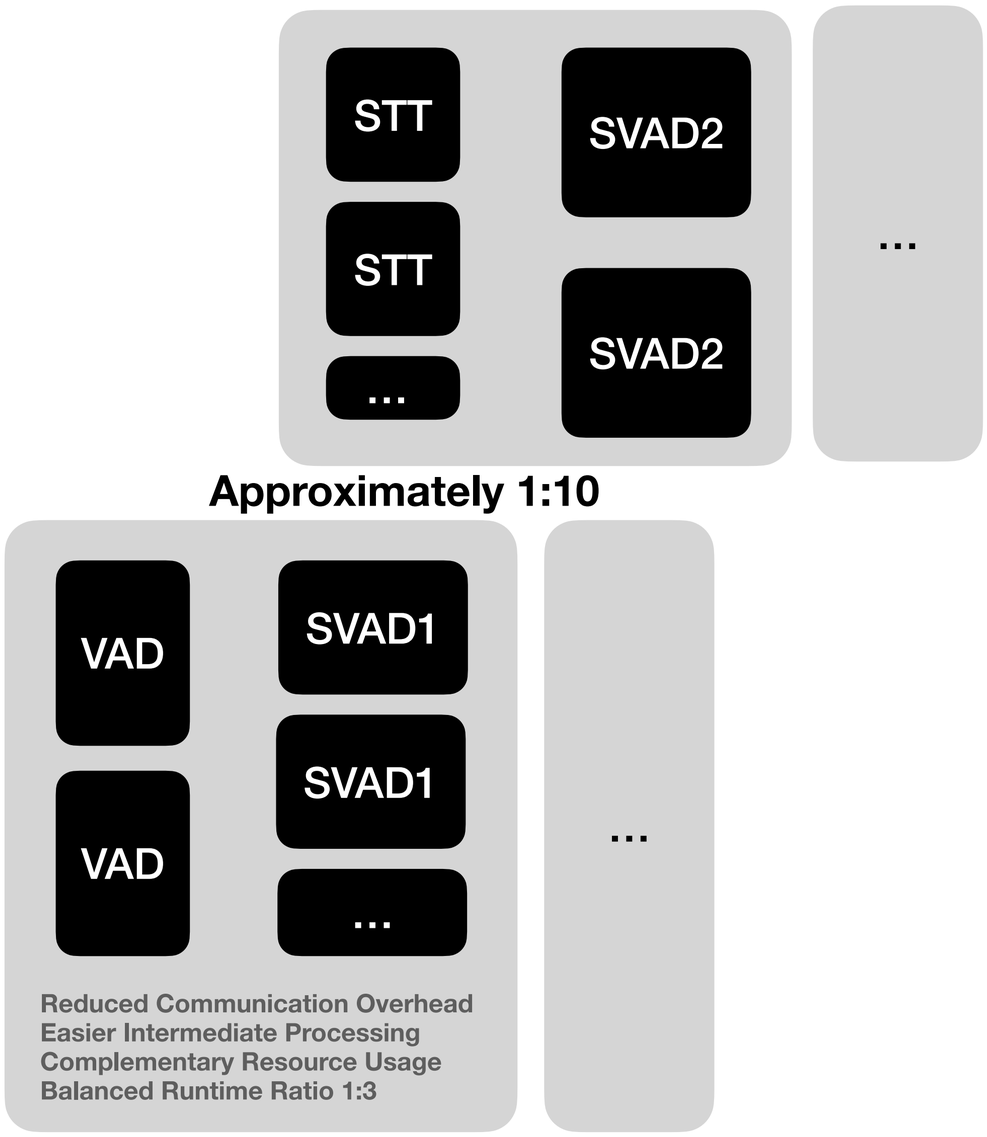
Distilhubert (excluding classification head)

GPU	Batch Size	Audio Duration	Approx. VRAM Usage
NVIDIA T4	1	1 sec	~150–200 MB
NVIDIA T4	1	10 sec	~250–300 MB
NVIDIA RTX 2080 Ti	8	10 sec	~1.0–1.2 GB
NVIDIA A100 (40GB)	32	10 sec	~2–2.5 GB

Distilbert-base-uncased (excluding classification head)

GPU	Batch Size	Max Sequence Length	Approx. VRAM Usage
NVIDIA T4	1	128	~200–250 MB
NVIDIA T4	1	512	~400–500 MB
NVIDIA RTX 2080 Ti	8	512	~1.2–1.5 GB
NVIDIA A100 (40GB)	32	512	~2–3 GB

DEPLOYMENT/COST



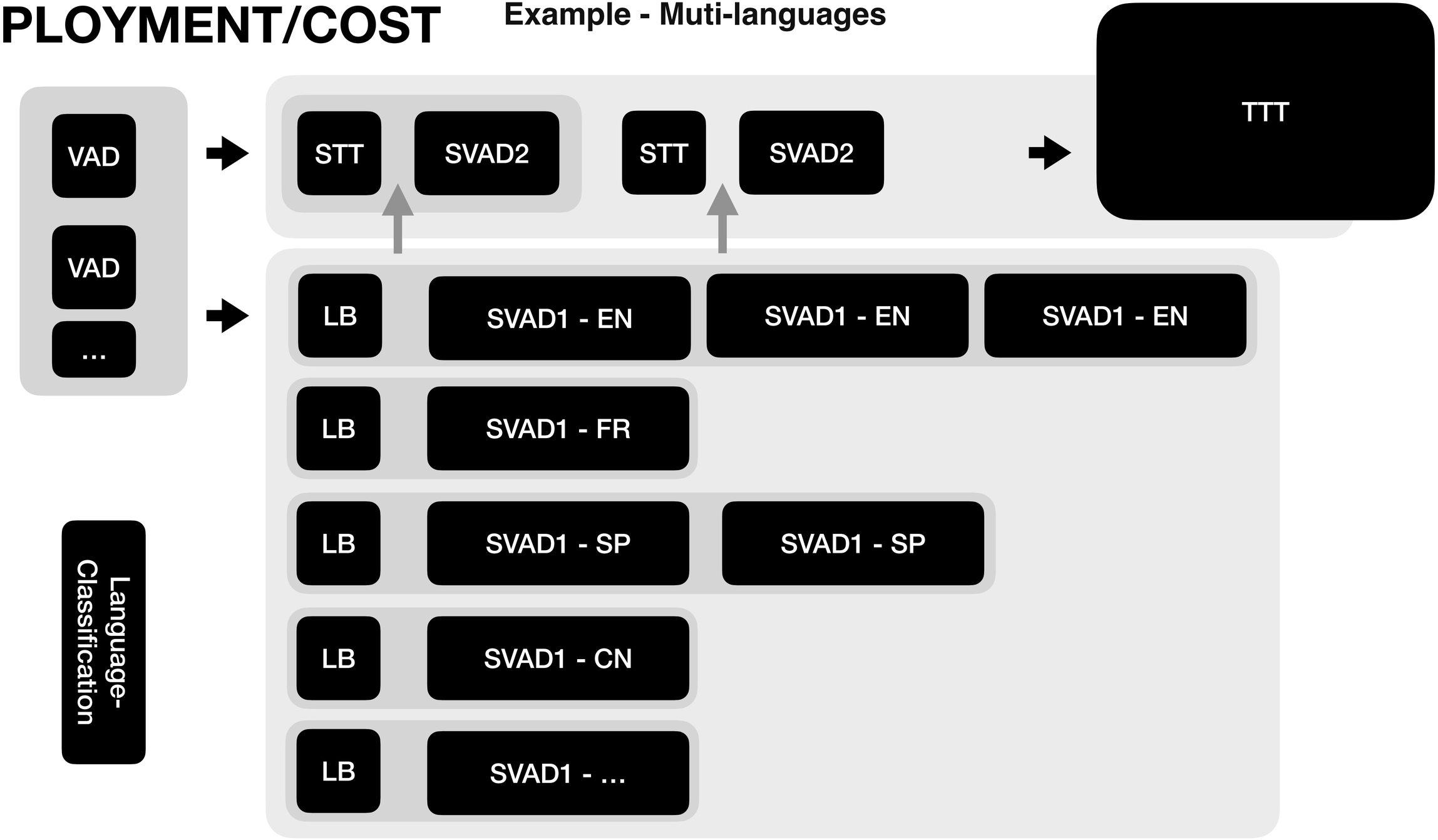
NCasT4_v3
~$0.90/hour (on-demand)

Parameter	Value
Audio duration	Real-time, 0.5s per segment
Segments per second per stream	2
Inference time per segment	2.5ms (optimized via batch size = 8)
GPU memory per group	700MB (1 VAD + 3 SVAD)
GPU total memory	16,000MB
Max number of groups on 1 GPU	⌊16,000 / 700⌋ = **22 groups**
Inference time per 0.5s segment	2.5 ms
Memory per VAD + 3×SVAD group	700 MB
Groups per T4 GPU (16GB)	22
Streams handled per group	200
Max real-time streams per instance	**~4400**

Ideally No System Overhead!
Ideally Perfect Loadbalancing!

DEPLOYMENT/COST

Example - Muti-languages



CONCERNS on POTENTIAL FAILURES

True Value \ Predicted Value	Backchannel	Interruption
Backchannel	CORRECT REFRENCE	Will user noticed? Can long silence be properly recovered? **Data-driven tuning based on business-specific logs**
Interruption	*“Oh,…,well,… I don’t….”* **Structural buffer-based adjustment**	CORRECT REFRENCE