ANDREW AND ERNA VITERBI
FACULTY OF
ELECTRICAL
ENGINEERING

SIPL

TECHNION
Israel Institute
of Technology

Few-Shot Learning Neural Network for Audio-Visual Speech Enhancement

Maya Rapaport

M.Sc. Seminar

Supervisor: **Prof. Israel Cohen**

Department of Electrical Engineering, Technion, Israel Institute of Technology

March 2021

Outline

Background & Motivation

Audio-Visual Speech Enhancement

Problem: Speaker Dependency

Proposed Algorithm

Few-Shot Learning

Problem: Dependency on the shots

Proposed Algorithm

Conclusions & Future Work

Research Contributions

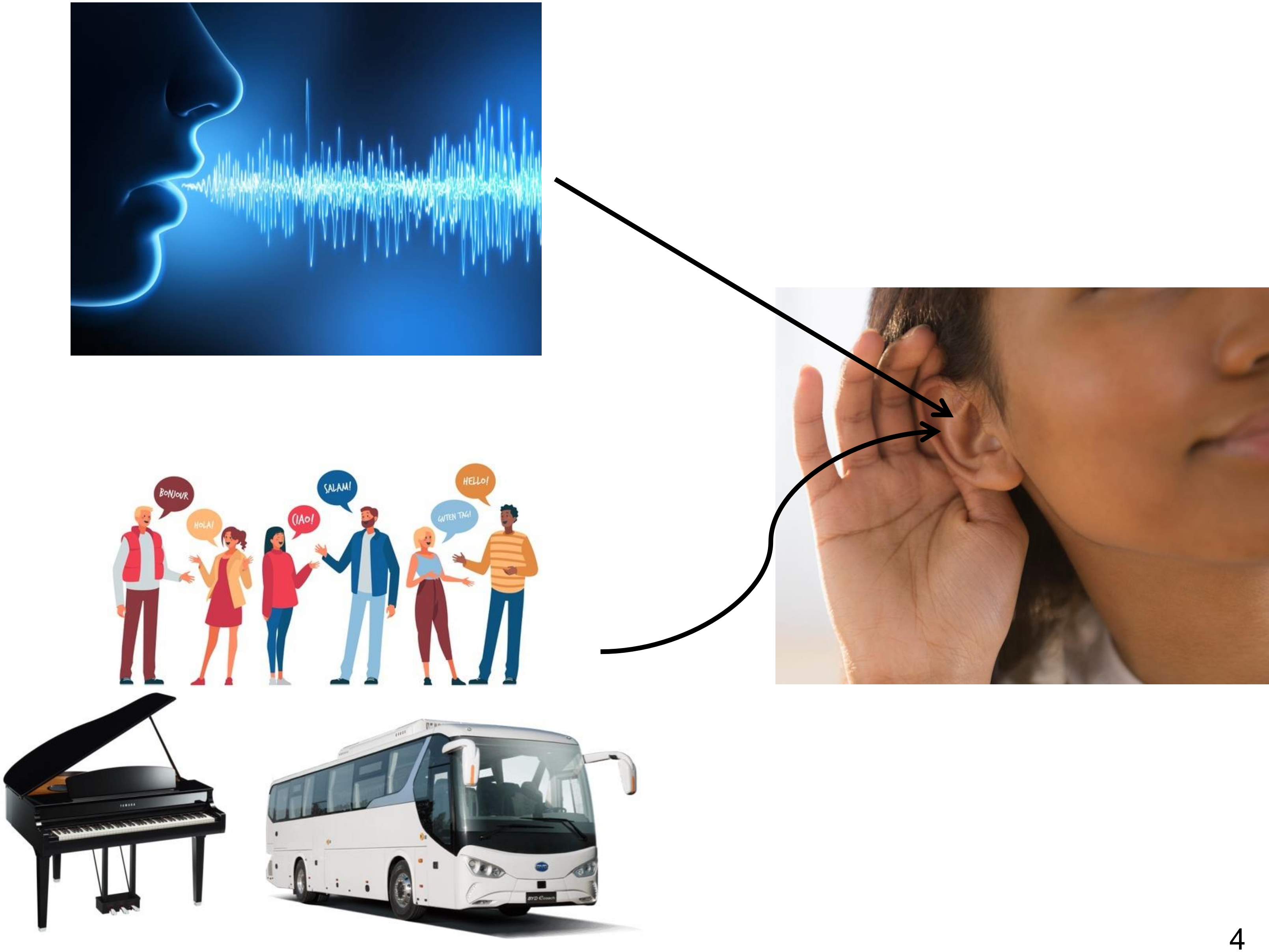
1. Overcoming speaker dependency for real-time mobile applications.

   Proposing
   Fast Adaptation Speech Enhancement (FASE) model,
   Inspired by few-shot learning methods

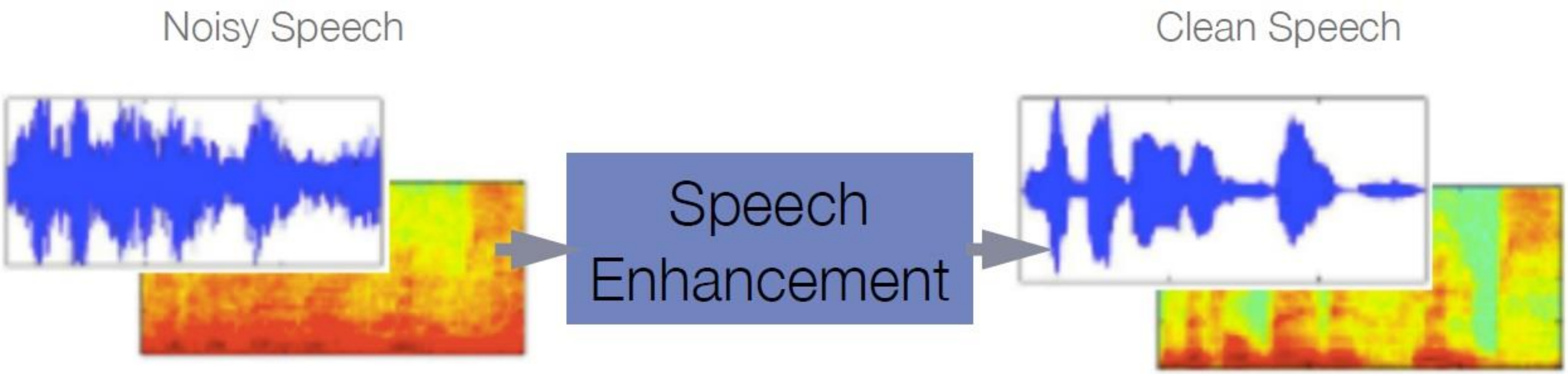
2. Extending few-shot learning to more shots.

   Proposing
   Novel algorithm to overcome few-shot learning limitations.
   Reduce dependency on the number of shots.

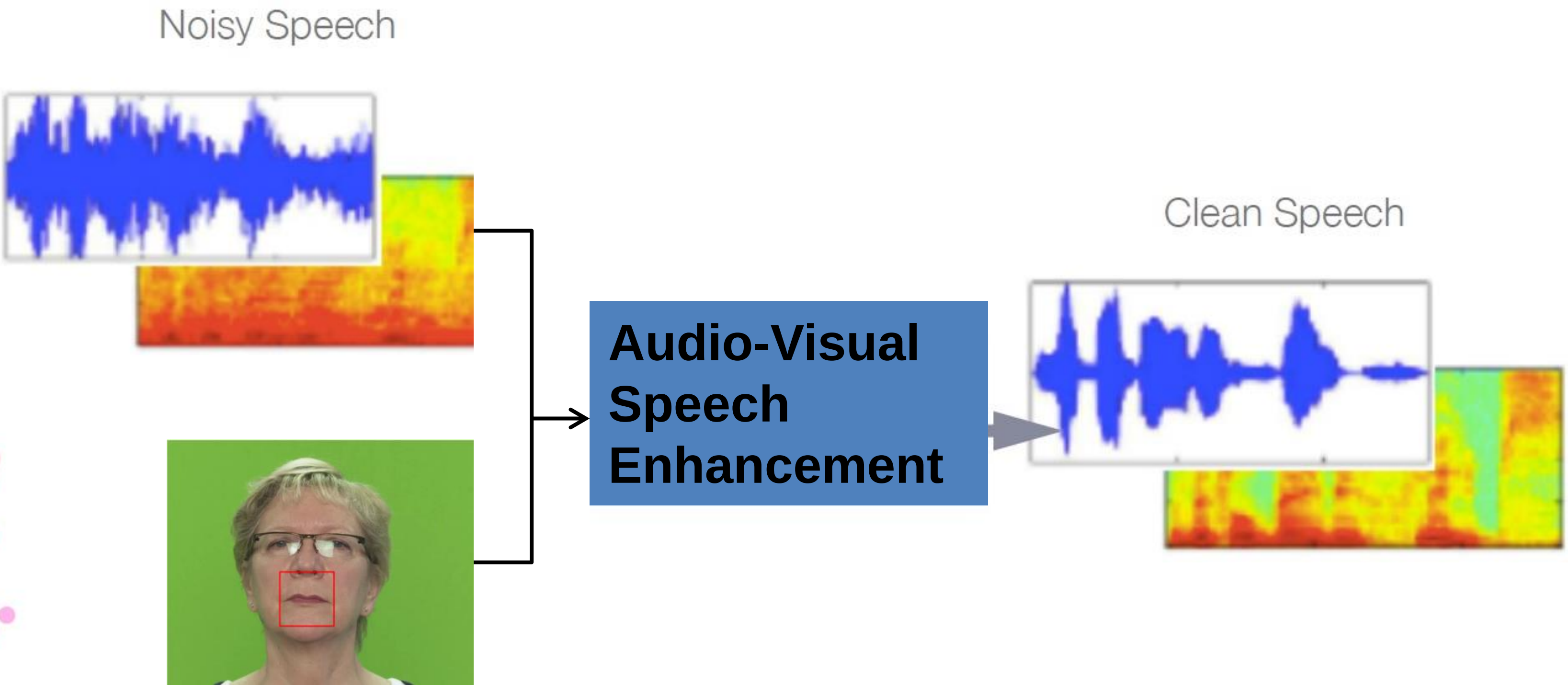
Background & Motivation



Speech Enhancement



Audio-Visual Speech Enhancement



Noisy Speech

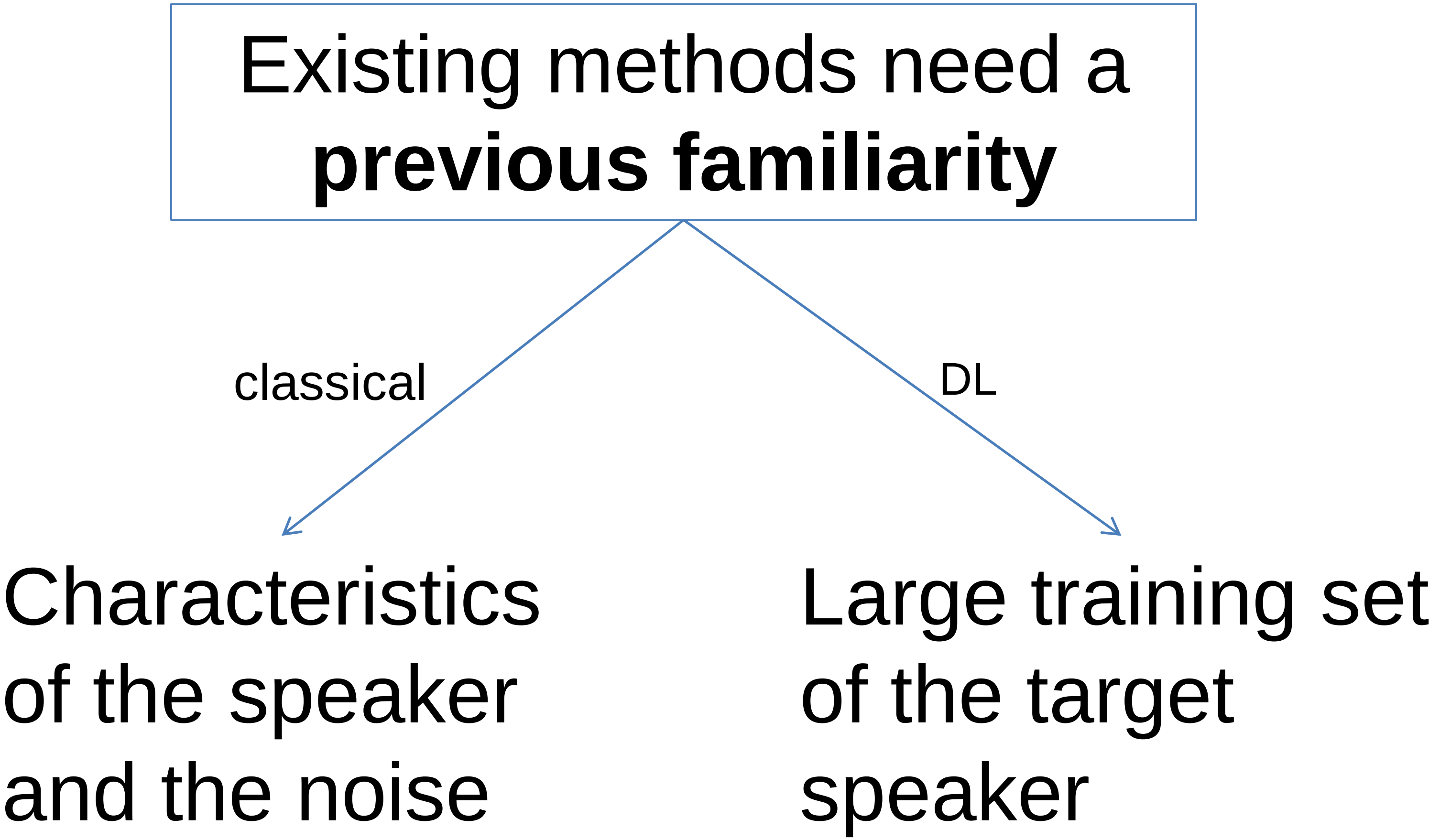
Audio-Visual Speech Enhancement

Clean Speech

Applications

Real-Time

The Problem – Speaker Dependency



Existing methods need a **previous familiarity**

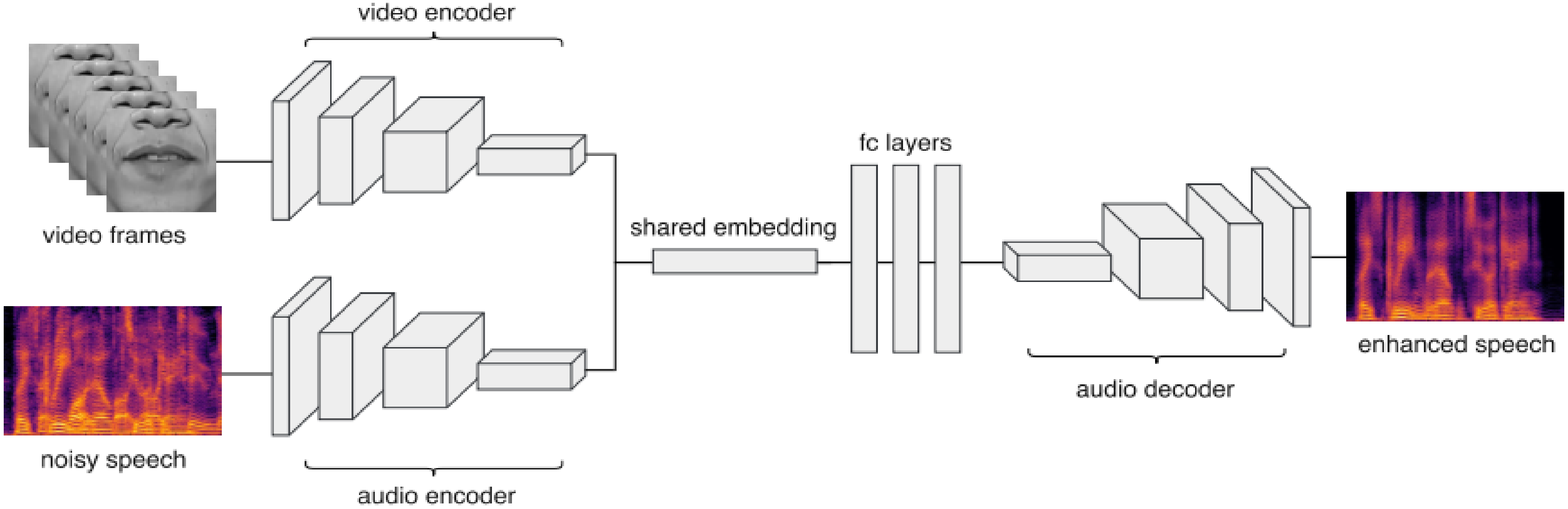
classical

DL

Characteristics of the speaker and the noise

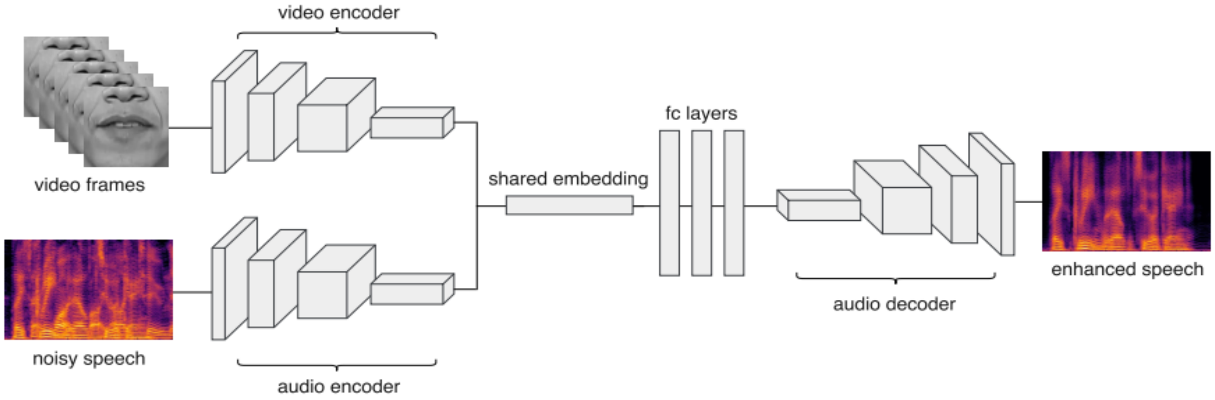
Large training set of the target speaker

The Problem – Speaker Dependency



A. Gabbay, A. Shamir, and S. Peleg, “Visual speech enhancement,” in Proc. Interspeech, 2018, pp. 1170–1174.

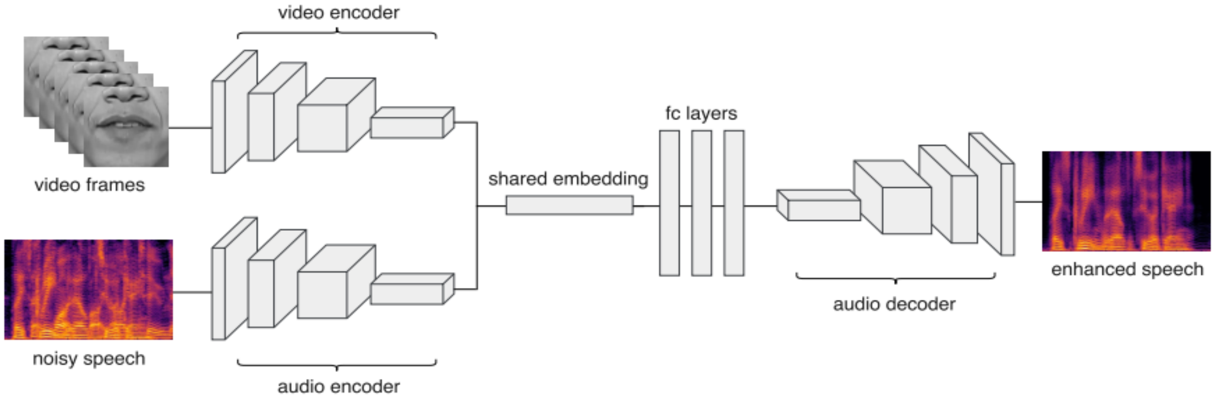
The Problem – Speaker Dependency



Experiment	Training		Test		Results (PESQ)
	Target	Noise	Target	Noise	
I	a	b	a	b	2.86
II	a	b	c	b	1.989
III	a	b	a	c	2.105

A. Gabbay, A. Shamir, and S. Peleg, “Visual speech enhancement,” in Proc. Interspeech, 2018, pp. 1170–1174.

The Problem – Speaker Dependency



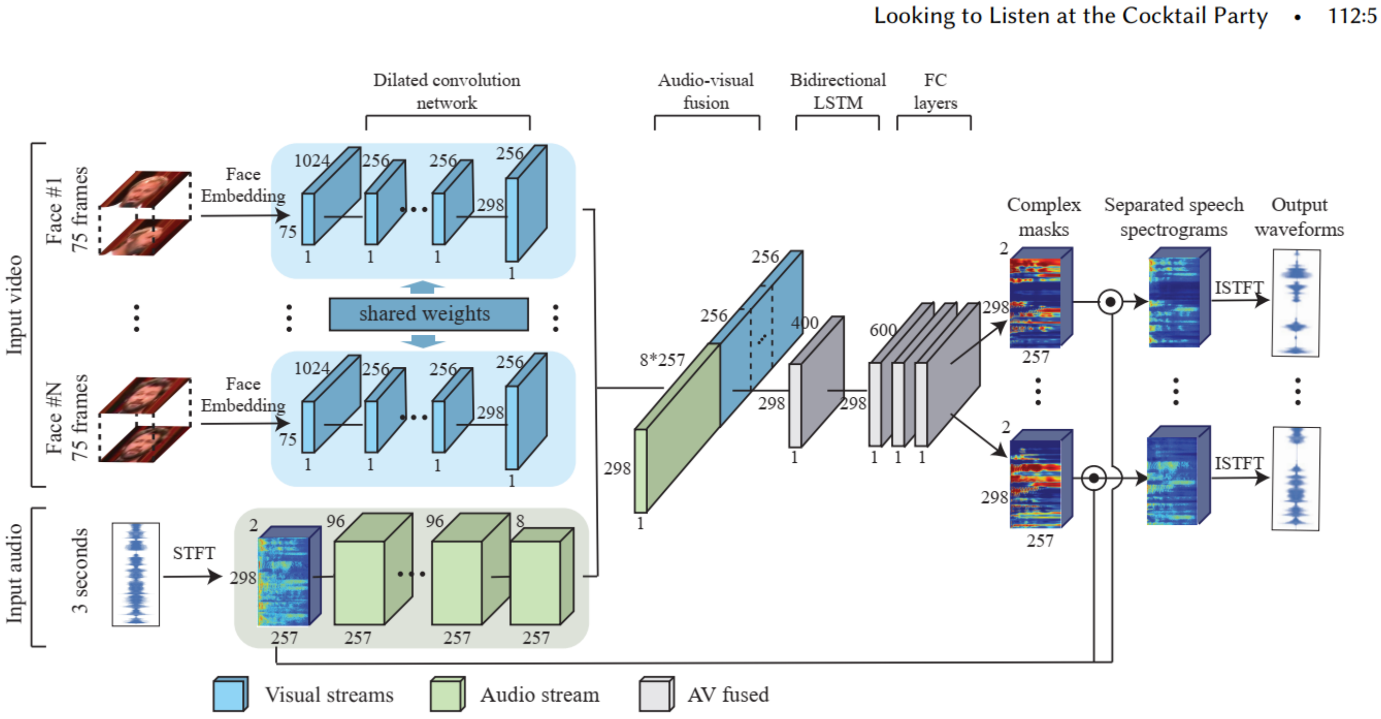
Suffers from Speaker Dependency

How to Avoid it?

A. Gabbay, A. Shamir, and S. Peleg, “Visual speech enhancement,” in Proc. Interspeech, 2018, pp. 1170–1174.

Avoiding Speaker Dependency

Very Large Dataset



High Computational Power

A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” *arXiv preprint arXiv:1804.03619*, 2018

Applications

~~Real-Time~~

Research Contributions

1. Overcoming speaker dependency for real-time mobile applications.

   Proposing
   Fast Adaptation Speech Enhancement (FASE) model,
   Inspired by few-shot learning methods

2. Extending few-shot learning to more shots.

   Proposing
   Novel algorithm to overcome few-shot learning limitations.
   Reduce dependency on the number of shots.

Research Contributions

1. Overcoming speaker dependency for real-time mobile applications.

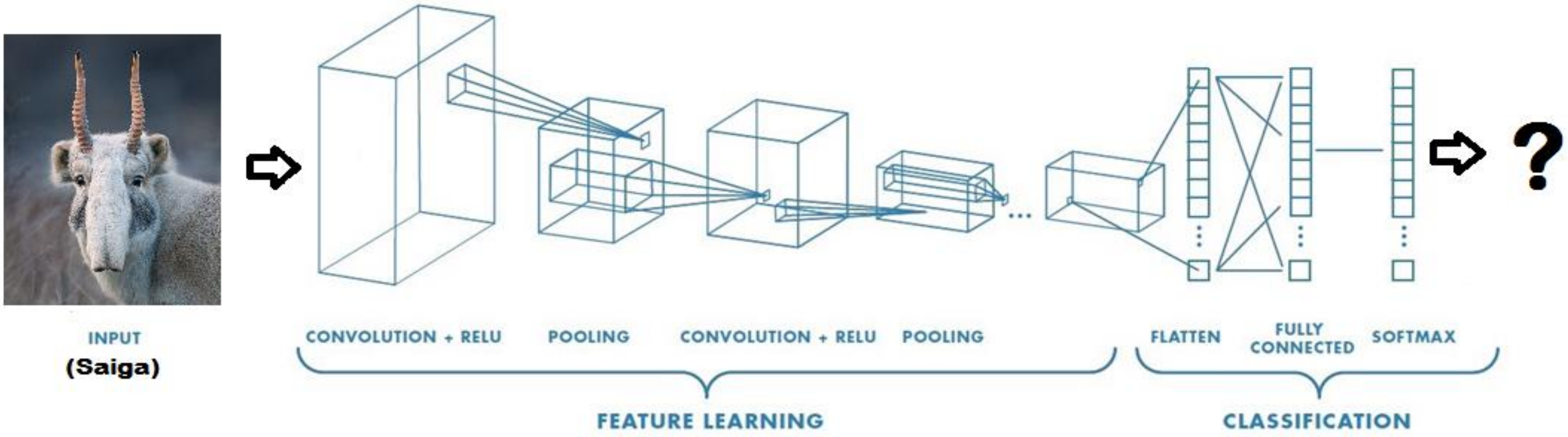
   Proposing
   Fast Adaptation Speech Enhancement (FASE) model,
   Inspired by few-shot learning methods

2. Extending few-shot learning to more shots.

   Proposing
   Novel algorithm to overcome few-shot learning limitations.
   Reduce dependency on the number of shots.

Few-Shot Learning

Image Classification



Not enough data to learn all classes!

Goal:

- Efficiently learn ***novel*** categories
- Not forget the ***base*** categories

n-way k-shot

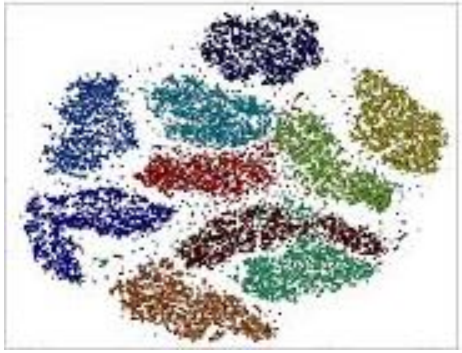
n = 1, 5
k = 1, 5

Very few compared to 600!

Main Approaches

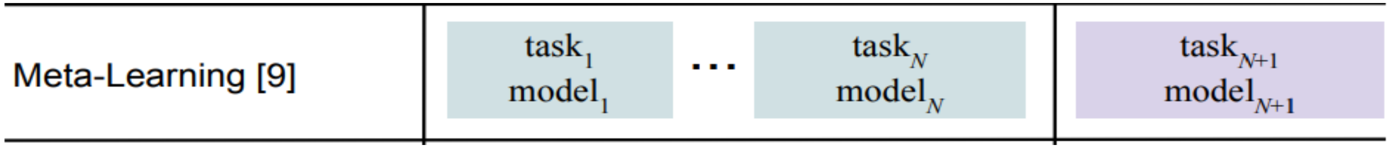
Metric-Learning:

Learn feature representations that preserve the class neighborhood structure.
→ Intra-class similarity and inter-class dissimilarity.



Meta-Learning (Learning to Learn):

Generate parameter updates that will optimize the classification performance of a learner model on a task.



Meta-Learning [9]

task_1 model_1 … task_N model_N | task_{N+1} model_{N+1}

Research Contributions

1. Overcoming speaker dependency for real-time mobile applications.

   Proposing

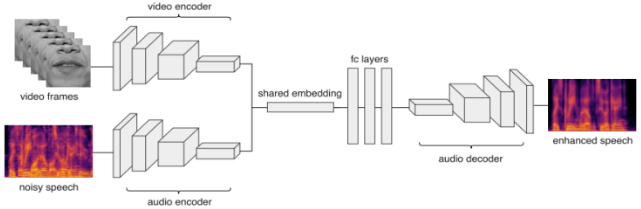
   Fast Adaptation Speech Enhancement (FASE) model,
   Inspired by few-shot learning methods

2. Extending few-shot learning to more shots.

   Proposing

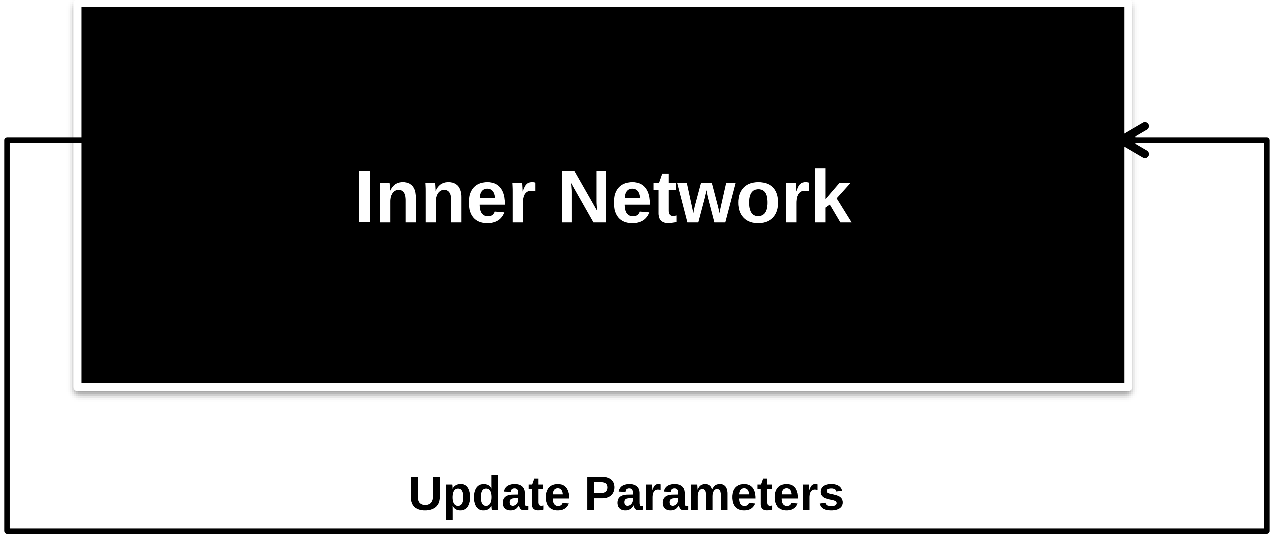
   Novel algorithm to overcome few-shot learning limitations.
   Reduce dependency on the number of shots.

FASE: Fast Adaptation Speech Enhancement



A. Gabbay, A. Shamir, and S. Peleg, “Visual speech enhancement,” in Proc. Interspeech, 2018, pp. 1170–1174.

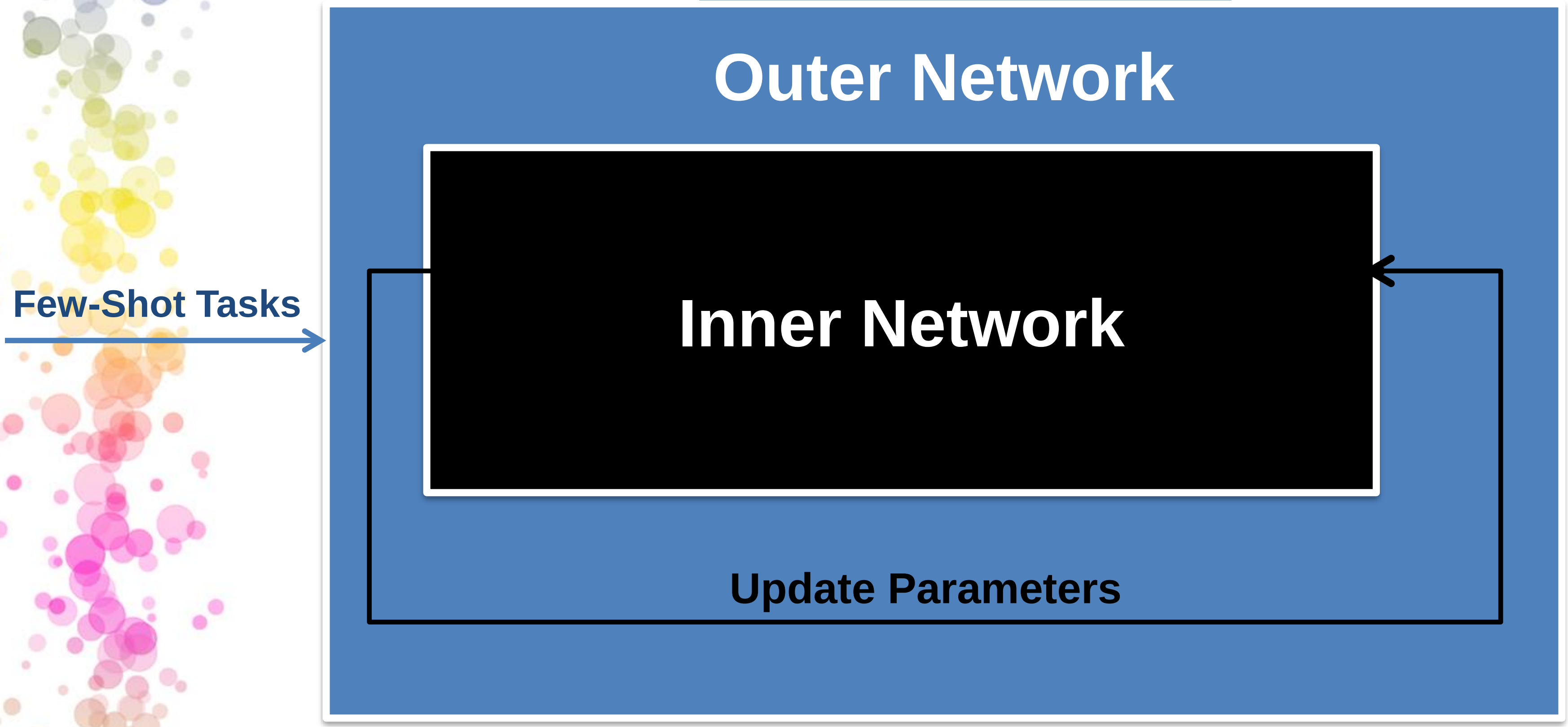
FASE: Fast Adaptation Speech Enhancement


Inner Network

Update Parameters

FASE: Fast Adaptation Speech Enhancement

FASE-baseline



Outer Network

Few-Shot Tasks

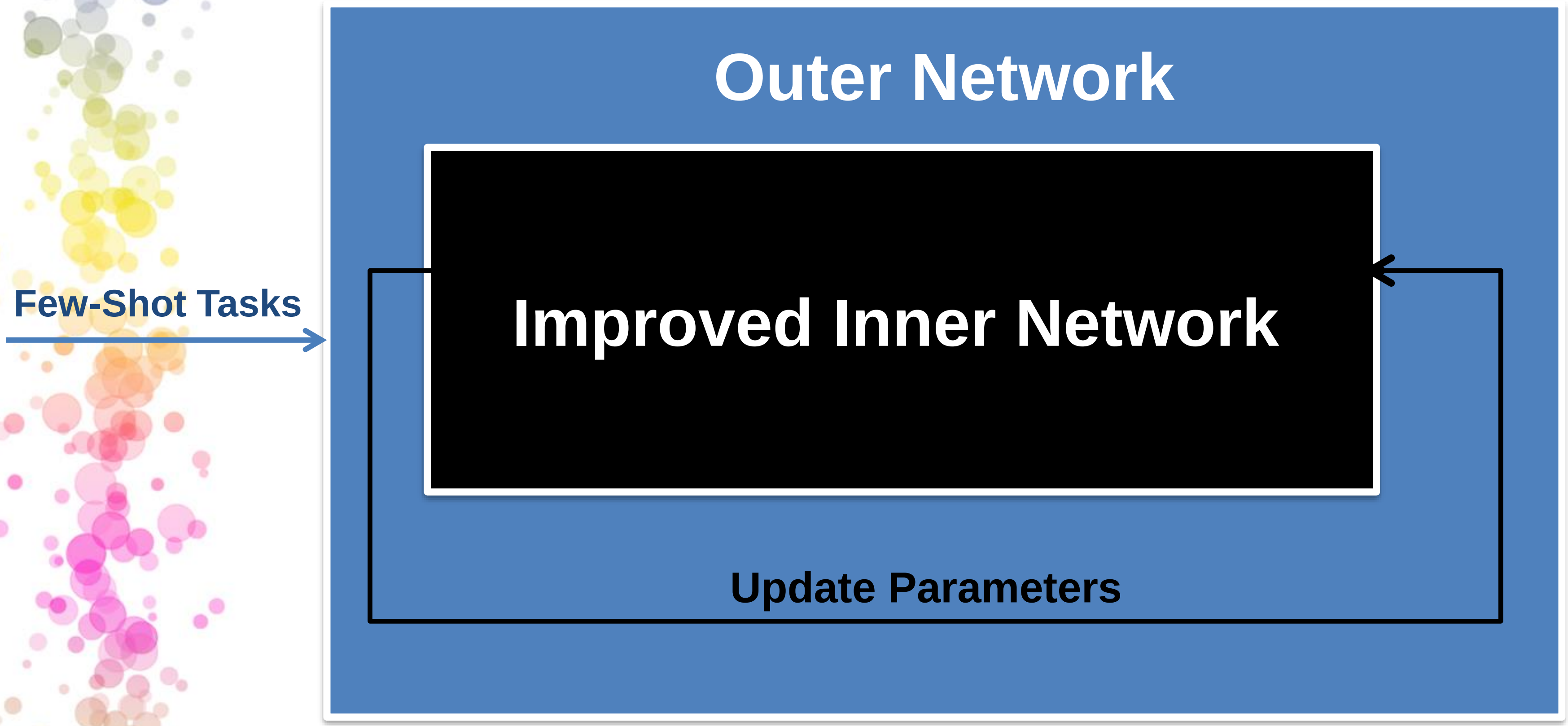
Inner Network

Update Parameters

Learns how to adapt to new speakers

FASE: Fast Adaptation Speech Enhancement

FASE-opt



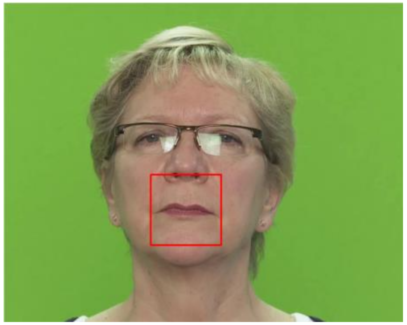
Faster adaptation and convergence without overfitting

FASE: Fast Adaptation Speech Enhancement

FASE Training Algorithm:

```
Choose number of shots $k$
Create a random pool of few-shot speech enhancement tasks.
For each few-shot task:
    Train the outer network (for M epochs):
        For each video:
            Train the inner network (for m epochs)
            Validate the inner network
        Update the outer network
    Validate the outer network
```

Experimental Settings



Dataset: TCD-TIMIT

Algorithms: Gabbai et al.
FASE-baseline
FASE-opt

Measurements: PESQ (**Q**uality)
STOI (**I**ntelligibility)

Number of Shots: $1 \le k \le 20$

Let’s Hear It ☺

Most challenging: One-Shot Learning (*k* = 1)

Base

Mixture

Gabbai et al.

FASE-baseline

FASE-opt

Novel

Mixture

Gabbai et al.

FASE-baseline

FASE-opt

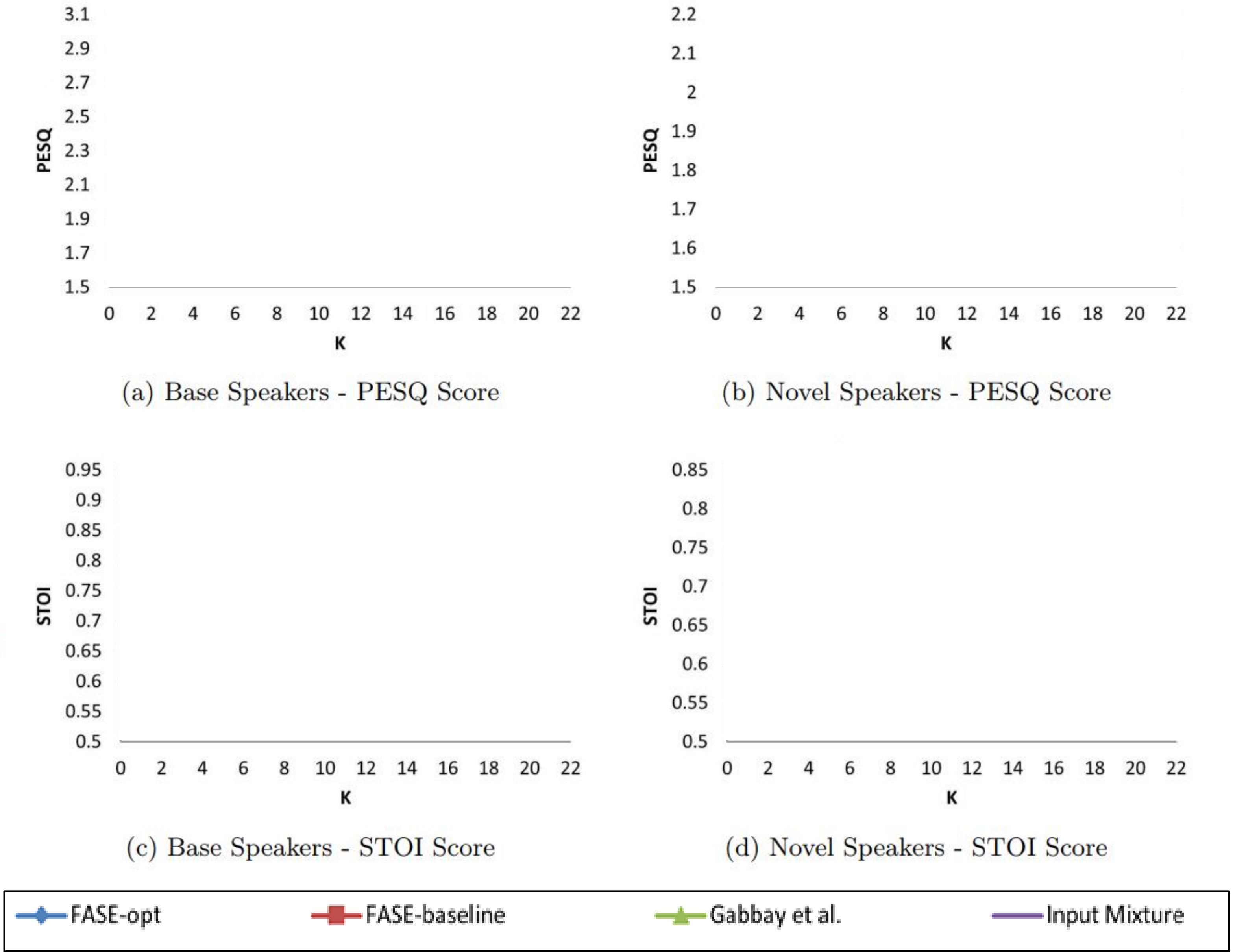
Mixture

Gabbai et al.

FASE-baseline

FASE-opt

FASE Model - Results



(a) Base Speakers - PESQ Score

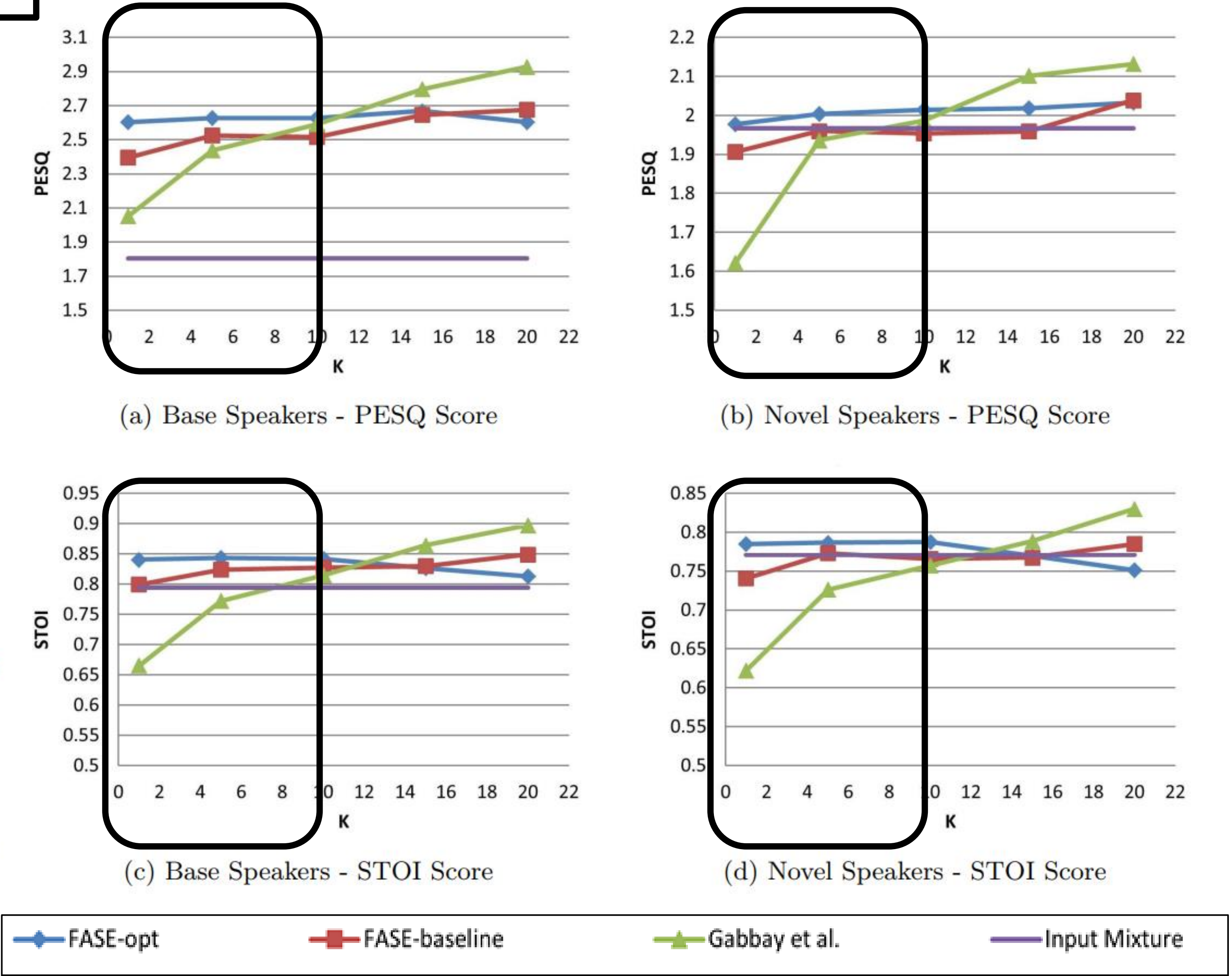
(b) Novel Speakers - PESQ Score

(c) Base Speakers - STOI Score

(d) Novel Speakers - STOI Score

“Standard” Few-Shot learning

FASE Model - Results



(a) Base Speakers - PESQ Score

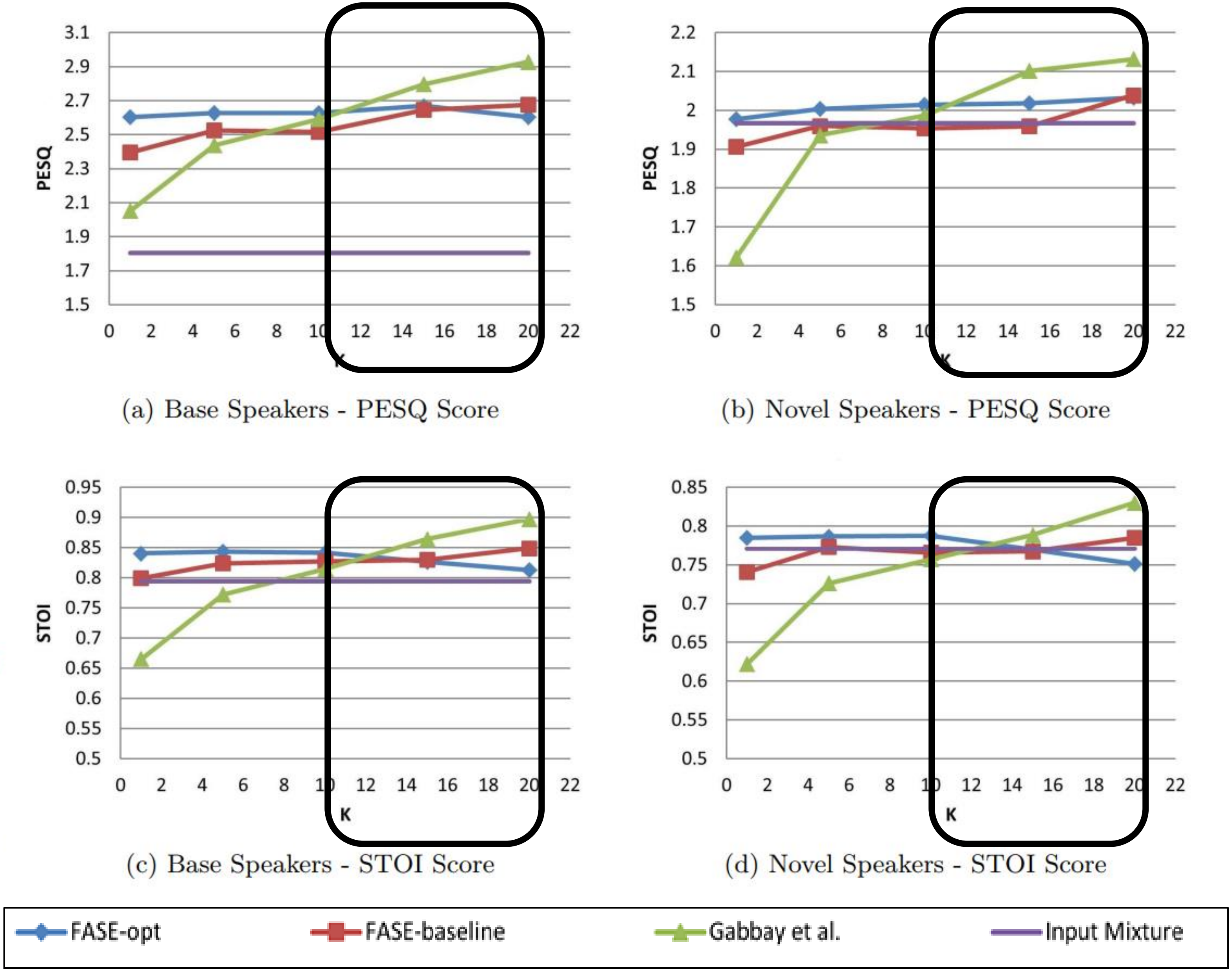
(b) Novel Speakers - PESQ Score

(c) Base Speakers - STOI Score

(d) Novel Speakers - STOI Score

Cases of More Shots

FASE Model - Results



(a) Base Speakers - PESQ Score

(b) Novel Speakers - PESQ Score

(c) Base Speakers - STOI Score

(d) Novel Speakers - STOI Score

Let’s Hear It

Largest gap: $k = 20$

Base	**Novel**
Mixture	Mixture
Gabbai et al.	Gabbai et al.
FASE-baseline	FASE-baseline
FASE-opt	FASE-opt

FASE Model - Summary

Fast Adaptation Speech Enhancement
Avoids Speaker dependency

“Standard” Few-Shot Learning

- ✓ Quality (PESQ)
- ✓ Intelligibility (STOI)
- ✓ Less Computational Power

Cases of More Shots

☹

↓

Let’s investigate... 🔍

Research Contributions

1. Overcoming speaker dependency for real-time mobile applications.

   <u>Proposing</u>
   Fast Adaptation Speech Enhancement (FASE) model,
   Inspired by few-shot learning methods

2. Extending few-shot learning to more shots.

   <u>Proposing</u>
   Novel algorithm to overcome few-shot learning limitations.
   Reduce dependency on the number of shots.

Research Contributions

1. Overcoming speaker dependency for real-time mobile applications.

   <u>Proposing</u>
   Fast Adaptation Speech Enhancement (FASE) model,
   Inspired by few-shot learning methods

2. Extending few-shot learning to more shots.

   <u>Proposing</u>
   Novel algorithm to overcome few-shot learning limitations.
   Reduce dependency on the number of shots.

Research Contributions

1. Overcoming speaker dependency for real-time mobile applications.

   Proposing
   Fast Adaptation Speech Enhancement (FASE) model,
   Inspired by few-shot learning methods

2. Extending few-shot learning to more shots.

   Proposing
   Novel algorithm to overcome few-shot learning limitations.
   Reduce dependency on the number of shots.

Few-Shot Learning Limitations

- Customized for $k = 1$, $k = 5$
- In reality – No guarantee
- From few-shots to more shots:
  - Exploring limitations
  - Proposed algorithm
  - Results (proof-of-concept)
  - Conclusions & Future work
- Gidaris et al. *

* S. Gidaris and N. Komodakis, “Dynamic few-shot visual learning without forgetting,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4367–4375.

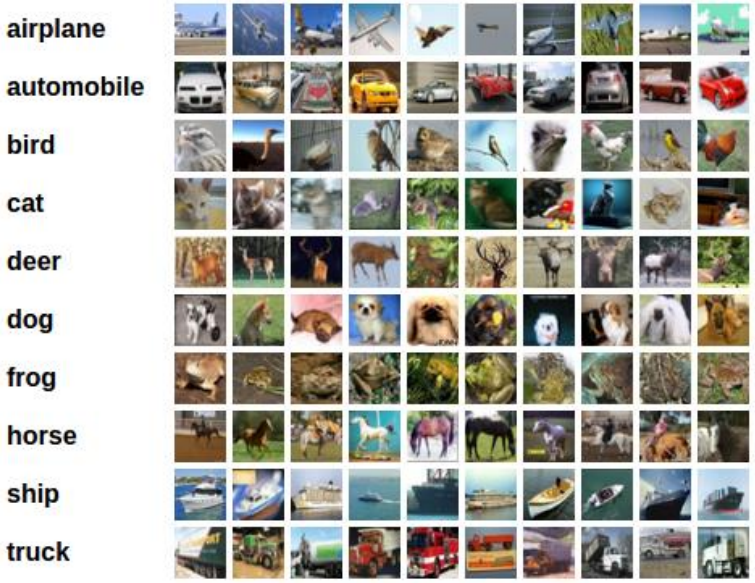
Experimental Settings

Dataset: mini-ImageNet

Algorithms: Gidaris et al.
Later – Our solution

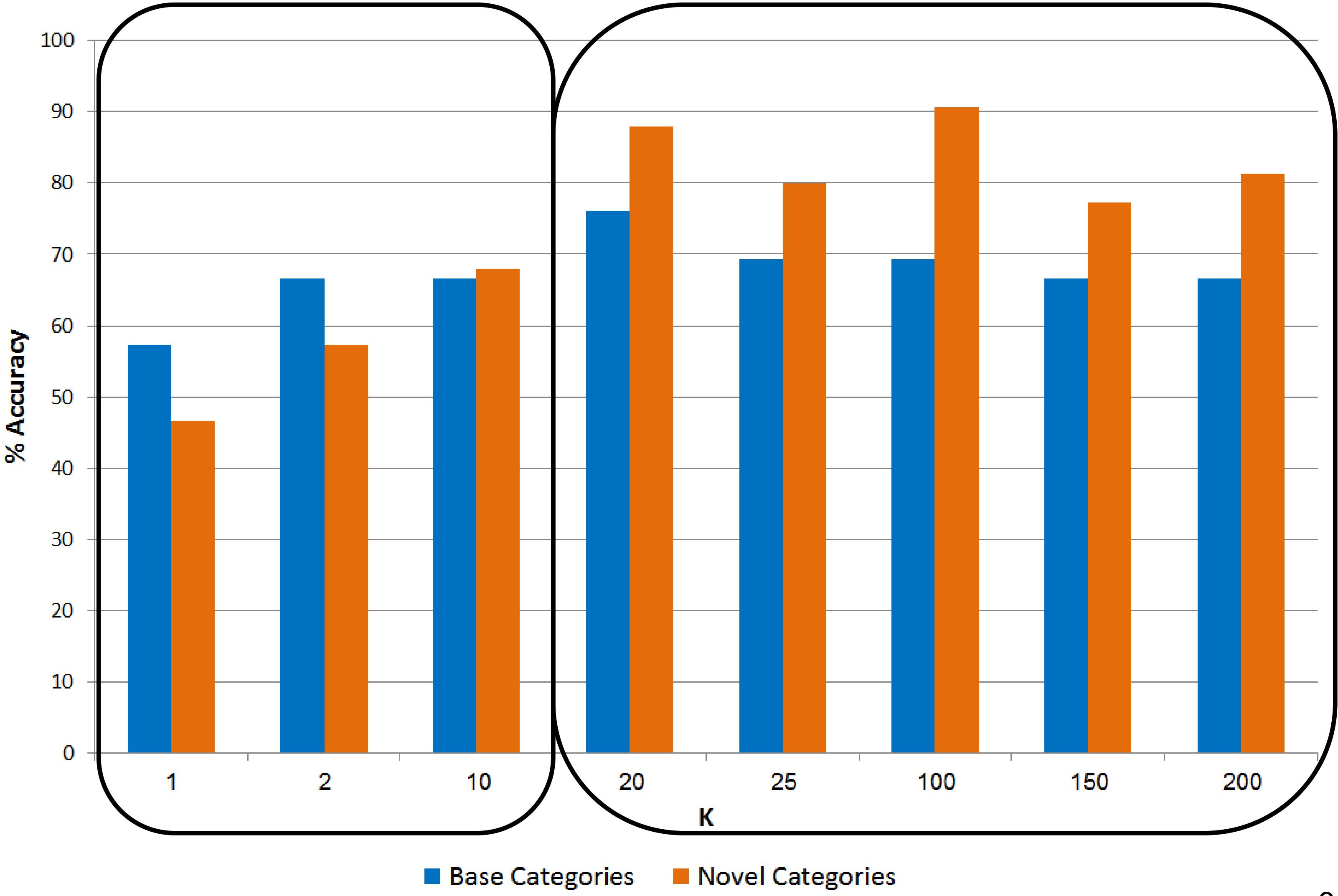
Measurements: % Accuracy

Number of Shots: $1 \le k \le 200$



Few-Shot Learning Limitations

STD along all cases: 9.25



From Few-Shot to More Shots

Proposed algorithm:

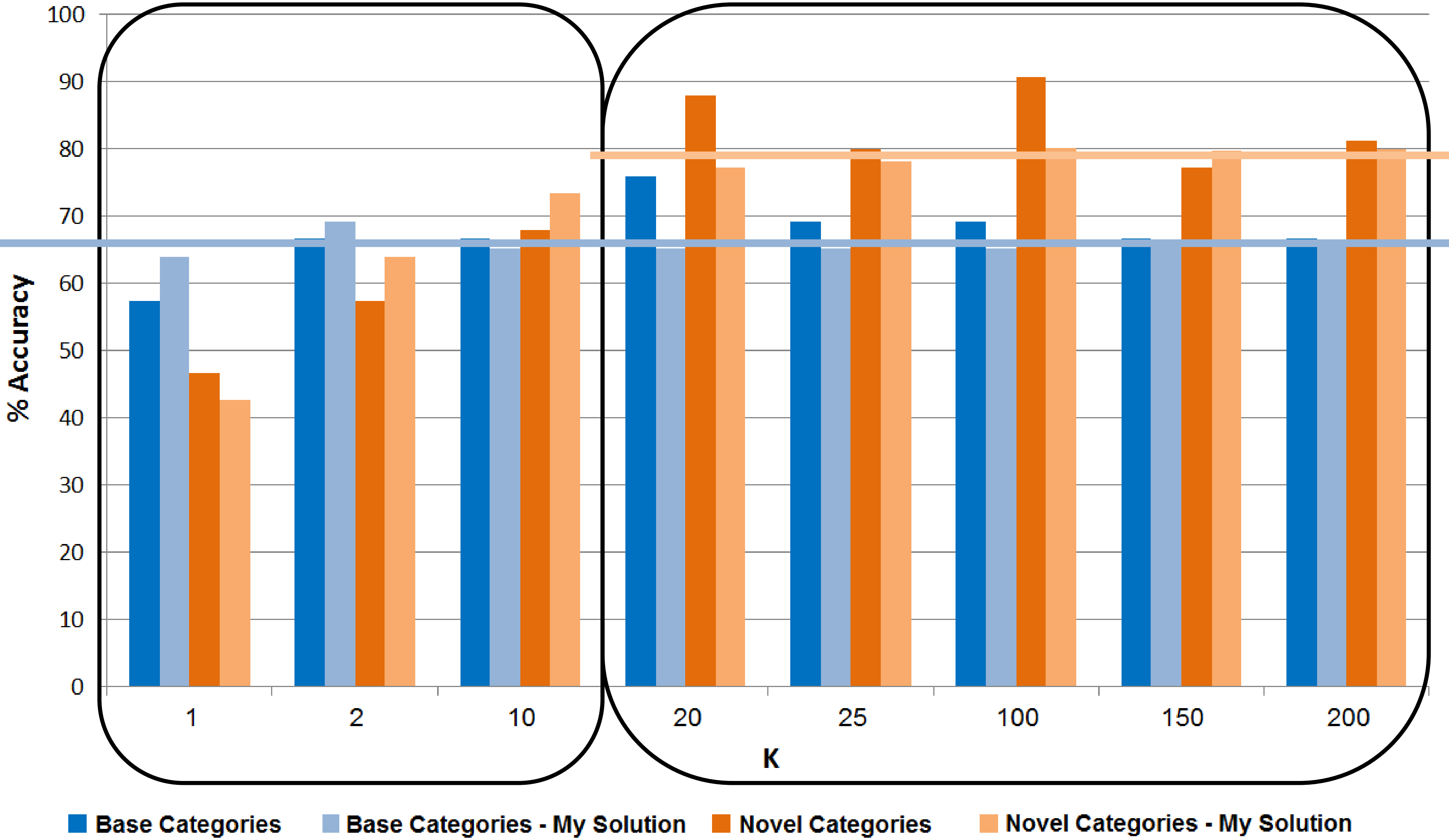
Train the feature extractor of Gidaris et al.
Divide the base feature vectors evenly into *N* spaces.
Continue as Gidaris et al.
For each image:
   Calculate matching in all *N* spaces
   Choose the best category by weighted calculation

Proof-of-concept ☺

Dividing method = Random
Training iterations = 20

Results

STD along all cases: 6.5 (instead of 9.25)



More Shots - Summary

Results:

- ✓ Smaller base-novel trade-off
- ✓ More shots: Improved stability
- ✓ Less dependency on *k*
- ✓ **The method is general**
  Can be used in other few-shot tasks

Future work:

- Optimize training.
- Explore the separation method into spaces.
- Find the ideal number of spaces.

Research Contributions

1. Overcoming speaker dependency for real-time mobile applications.

   - ✓ Quality (PESQ)
   - ✓ Intelligibility (STOI)
   - ✓ Less computational power

2. Extending few-shot learning to more shots.

   - ✓ Improved stability
   - ✓ Smaller base-novel trade-off
   - ✓ General method

Future Research

- Improving few-shot learning model for more shots.
  → Training, space separation, number of spaces

- Improving results of FASE model with few shots.
  → Larger dataset, training configurations

- Improving results of FASE model with more shots.
  → Deploying the proposed **general solution** in audio-visual speech enhancement
  → Learn a **filter** rather than the spectrogram itself. (inability to create sounds that do not exist in the original soundtrack)

Thank You ☺

Special thanks to Israel Cohen
For all the support and professional guidance