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

Maya Rapaport

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

Research Thesis

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and
Computer Engineering

Maya Rapaport

Submitted to the Senate
of the Technion — Israel Institute of Technology
Iyar 5781 Haifa May 2021

This research was carried out under the supervision of Prof. Israel Cohen, in the Faculty of Electrical and Computer Engineering.

Some results in this thesis have been submitted as manuscript by the author and supervisor to IEEE/ACM Transactions on Audio, Speech, and Language Processing, January 2021.

Acknowledgements

I would like to thank my supervisor Prof. Israel Cohen.

Contents

List of Figures

List of Tables

Abstract

Audio-Visual speech enhancement is the task in which given a noisy audio signal and the corresponding speaker video frames, the model would produce an enhanced audio signal containing only the target speaker's voice, whereas the rest of the speakers and background noise would be suppressed. As opposed to humans, it is a significant challenge for audio-visual speech enhancement models to isolate and enhance a target speaker without having any previous familiarity with the speakers in the video. This problem is known as speaker dependency. It prevents speech enhancement models from performing in real-time applications, where a previous familiarity cannot be guaranteed.

In this thesis we look at the realistic problem of having only a small number of training samples of the target speaker. We propose a fast adaptation speech enhancement (FASE) model, a state-of-the-art audio-visual speech enhancement neural network. Our model is inspired by methods which were originally developed for the task of few-shot learning in image classification, and specifically the meta-learning approach. The model comprises a deep encoder-decoder architecture, which is trained by an outer network model for fast adaptation to enhance new speakers. Moreover, we propose an improved encoder-decoder architecture.

We show that when there is only a few samples of the target speaker, FASE model outperforms previous audio-visual speech enhancement models, both in speech quality and intelligibility of all speakers. The model also demonstrates an improvement in computational performance, which implies its potential application for mobile systems, such as smart hearing aids. In cases of more samples of the target speaker our model exhibits limitations, similarly to other few-shot learning methods.

Few-shot learning algorithms are customized and designed for the cases of very few training samples. This is a disadvantage, since we would like the performance to improve when introducing the network with more training data. Therefore, we investigate the limitations of few-shot learning algorithms when coping with more shots in the task of image classification. We observe an instability of the performance due to the dependency of the results on the number of training samples. We propose an algorithm to overcome this disadvantage, using the feature vectors extracted from the images. Our algorithm manages to decrease the undesirable dependency and achieve a better stability. Since the proposed algorithm is general and not specifically designed for images, it has a great potential to fit other few-shot learning tasks. We believe that it can be also used for an improved independent audio-visual speech enhancement model.

Glossary

Abbreviations

ASR	Automatic Speech Recognition
FASE	Fast Adaptation Speech Enhancement
GAN	Generative Adversarial Network
MSE	Mean-Squared Error
MUSHRA	Multiple Stimuli with Hidden Reference and Anchor
PESQ	Perceptual Evaluation of Speech Quality
STOI	Short-Time Objective Intelligibility
STFT	Short-Time Fourier Transform

Notations

$Att(\cdot,\cdot)$	attention kernel
$c(\cdot\|W^*)$	classifier
$G(\cdot,\cdot\|\phi)$	classification weight generator
k	number of novel training samples (number of shots)
K_{base}	number of base categories
n	number of novel classes
p	classification probability vector
$p(T)$	distribution of classification tasks
s_k	classification score
S_{n_k}	score to the category k in each space n
T	classification task
T_{test}	test split of classification tasks
T_{train}	train split of classification tasks
T^{new}	unseen task
T^{new}_{test}	set of unseen data points
$W^{'}$	classification weight vector
$W^{'}_{att}$	The attention-based classification weight vector
z	extracted feature vector
ϕ_{avg}	learnable weight vector
ϕ_q	cosine similarity function

Chapter 1

Introduction

1.1 Background and Motivation

In real-life situations it is almost unavoidable to capture audio signals without any background noise. The noise may be formed both by the surrounding environment and by the very system recording the signal. While humans have the ability to mentally "mute" background noise and concentrate on a specific speaker, the ability of automatic speech separation is a significant challenge for computers. The noise degrades the signal quality and intelligibility before being stored, transmitted or played.

Speech enhancement models get a noisy audio signal input and produce an enhanced audio signal containing only the target speaker's voice, whereas the rest of the speakers and background noise are suppressed. Audio-visual speech enhancement correlates the audio with the visual features of the target speaker in order to improve speech enhancement results [1]. Intuitively, visual features, such as lips movement, should correlate with the speech and can help identify which parts of the audio correspond to that speaker.

The problem of speech enhancement, with or without a visual input, is very challenging and has been extensively researched over the past few decades. Recently, there is an increased interest in using neural networks with audio-visual inputs, outperforming state-of-the-art methods for speech separation and enhancement. The visual aspect of speech, which is essentially unaffected by the acoustic environment, can be efficiently fused by exploiting the flexibility of data-driven approaches, specifically deep learning methods.

However, most of the speech enhancement models, and neural networks in particular, suffer from a problem called speaker dependency. In order to enhance a target speaker, such models must have previous familiarity with the speakers in the video [2], [3]. Practically, the training set should comprise of clean samples of the target speaker. Ideally, those models must have an understanding of the characteristics of the target speech, and preferably also on the background noise. This dependency results in a challenging problem which prevents speech enhancement algorithms from running in real-time applications, where a previous familiarity cannot be guaranteed.

Speech enhancement models are required in many real-time applications:

- **Hearing Aids.** Speech enhancement can help processing the noisy signal to reduce the noise prior to amplifying it for the hearing impaired.
- **Automatic Speech Recognition (ASR).** Speech enhancement pre-processing can address the significant deterioration in performance for automatic recognition in noisy environments. For example, audio signals from mobile phones in a crowded environment can be enhanced before being analyzed for recognition.
- **Telecommunications.** Voice communication suffers from the background noise present at the transmitting end, which can be processed to reduce the noise and improve speech quality prior to being played at the receiver end.
- **Automatic Captioning.** Handling overlapping speakers is a known challenge for automatic captioning systems. Separating the audio to distinct sources and denoising could help presenting more accurate and easy-to-read captions.
- Additional applications include virtual assistants, personalized advertisement, and a variety of applications for intelligence corps, health and wellness.

1.2 Overview of the Thesis

In this work we address the problem of speaker dependency, and concentrate mainly on real-time low-power applications. We propose a fast adaptation speech enhancement (FASE) network for overcoming the problem of speaker dependency. The proposed model is inspired by meta-learning approaches to the problem of few-shot learning in image classification tasks [4]. Few-shot learning algorithms are used to efficiently learn *novel* categories from only a few training samples, and nonetheless keep high performance on the *base* categories which the model was trained on. We observe that the target speaker can be referred to as a category which has only a few data samples for training the model. Thus, few-shot learning approaches can be used for avoiding speaker dependency.

While the advantages of few-shot learning models are substantial, the drawbacks should not be overlooked. The challenge of few-shot learning models is to avoid overfitting due to integration of the prior knowledge, comprising of a large number of parameters, with a small amount of new information [5]. In this work we extensively investigate the limitations of few-shot learning algorithms, both in the proposed FASE algorithm and in the original task of image classification. We observe a trade-off between the performance of *base* and *novel* categories, which is affected by the number of *novel* training samples (the number of shots). We propose a solution to overcome this drawback and achieve a better stability along different numbers of shots. The proposed solution is general and thus can be deployed in many few-shot learning tasks.

1.3 Research Contributions

The main contributions of our research are as follows:

- **Avoiding speaker dependency in audio-visual speech enhancement tasks.** We develop a fast adaptation speech enhancement neural network, which overcomes the problem of speaker dependency. The proposed model is inspired by methods of meta-learning in the task of few-shot image classification. Our model outperforms prior art in both quality and intelligibility measurements, examined on both the novel speakers (those with only few training samples) and for the base speakers (with a large representation in the training set). FASE model also demonstrates improvements in compute time. To the best of our knowledge, FASE model is the first meta-learning architecture to address the problem of speaker dependency in audio-visual speech enhancement, using few-shot learning approaches.
- **Extending few-shot learning methods to cope with more shots.** We start by examining few-shot learning methods in the task of image classification, and observe a base-novel trade-off which is dependent on the number of novel training samples. We propose a new approach to overcome this disadvantage. Our algorithm involves a relatively simple change in the space of the feature vectors extracted from the images by the network. We provide a proof-of-concept, which achieved better stability across different amounts of novel training samples. Our algorithm is general and therefore can be useful for other few-shot learning tasks, such as for avoiding speaker dependency.

1.4 Thesis Organization

The rest of the thesis is organized as follows: In Chapter 2, we introduce related background and approaches for both audio-visual speech enhancement and few-shot learning. In Chapter 3, we propose a fast adaptation speech enhancement network for avoiding speaker dependency in audio-visual speech enhancement tasks. We discuss its results, advantages and drawbacks. Chapter 4 further investigates the limitations of few-shot learning algorithms and suggests a possible solution. Then, we conclude the research and propose future work in Chapter 5.

Chapter 2

Background

As opposed to humans, most neural networks for audio-visual speech enhancement cannot automatically mute background noise and concentrate on the target speaker. They suffer from a problem known as speaker dependency. It implies that in order to enhance a target speaker in a video, such models must have previous familiarity with this speaker. The goal of our research is to decrease speaker dependency without constructing a massive network which requires large dataset and computational power.

Our solution aggregates methods of few-shot learning with the task of audio-visual speech enhancement. We look at new target speakers as few training samples and use meta-learning methods in order to quickly adapt the network to their new features. Then, we further analyze few-shot learning methods in the task of image classification and propose a solution to their drawbacks.

In this chapter we introduce some background and inspiring researches from all the relevant fields. In Section 2.1 we introduce the task of audio-visual speech enhancement, including some historical research background and main practices. In Section 2.2 we discuss the problem of speaker dependency, emphasizing why our goal is to avoid it. Section 2.3 introduces the problem of few-shot learning, originally tackled in the task of image classification, and presents existing approaches to cope with it.

2.1 Audio-Visual Speech Enhancement

Speech enhancement is the task whose purpose is to extract a target speech signal from a mixture of sounds generated by several sources. Traditionally, speech enhancement tasks have been tackled using signal processing techniques applied to the available acoustic signals. Over the past few decades the proposed methods included spectral subtraction techniques [6], [7], optimal filtering [8], [9], statistical-model-based algorithms [10], [11], subspace methods [12], and binary mask methods [13].

Since the visual aspect of speech is essentially unaffected by the acoustic environment, visual information from the target speaker, such as lip movements and facial expressions, has been recently introduced to speech enhancement systems, forming the research field of audio-

visual speech enhancement. An overview of key methodologies for classical audio-visual source separation methods can be found in [14].

However, the vast majority of classical audio-only and audio-visual speech enhancement algorithms require an estimate of the noise spectrum, since it has a large impact on the enhanced signal. If the estimate is too low then residuals of the noise might remain. On the other hand, when the estimate of the noise is too high, the enhanced speech is usually distorted. Moreover, there is a trade-off between noise reduction and speech distortion, as suppressing the noise usually introduces perceptible distortion to the speech signal. Thus, there is a compromise between the quality and intelligibility of the enhanced signal [3], [15].

More recently, machine and deep learning algorithms have been proposed for speech enhancement tasks, surpasses former algorithms [16], mainly thanks to their ability to learn a nonlinear mapping and to recognize a pattern without making explicit assumptions about the background noise or speech. Deep learning audio-only methods are widely used for speech denoising as well as for speech separation tasks. Wang and Chen [17] give a comprehensive overview of those methods. Lu et al. [18] propose deep auto-encoder architecture for the task of denoising a speech signal. Their model predicts a mel-scale spectrogram which represents the clean speech. Deep neural networks have been also trained to differentiate between speech characteristics of different sources in the tasks of speech recognition and separation. Speech characteristics, such as pitches, spectral bands and chirps, are unique and thus can differentiate between speakers [19], [20]. Pascual et al. [21] use generative adversarial networks (GANs) at the waveform level. It is worth mentioning that audio-only approaches achieve lower performance when separating similar human voices, e.g. same-gender mixtures [19].

Neural networks have been also applied in many tasks of audio-visual speech processing. Ngiam et al. [22] demonstrate cross modality feature learning. They show that better features can be learned if both audio and video are present during feature learning for one modality. Audio-visual inputs have also been used with multi-modal neural networks for tasks of lip sync [23], lip reading [17], and robust speech recognition [24]. In the task of speech enhancement, adding the visual signal as input not only improves the speech separation quality significantly in cases of mixed speech, but also associates the separated, clean speech tracks, with the visible speakers in the video [1]. Hand-crafted visual features are used in [25] in order to derive binary and soft masks for speaker separation. For audio-visual speech enhancement, Hou et al. [26] use a convolutional neural network model. They proposed using as input both a sequence of frames cropped to the speaker's lips region, and a spectrogram representing the noisy speech. Their output is a spectrogram representing the enhanced speech. A trained speech generation network was proposed by Gabbay et al. [27]. The video frames are used together with the spectrogram of the predicted speech, for constructing masks in order to separate the clean voice from the noisy input. The method presented by Afouras et al. [28] does not include the spectrograms as images but rather as temporal signals with the frequency bins as channels. Thus, allowing them to build a deeper network that trains fast despite the large number of parameters. Instead of directly predicting the clean magnitudes, their model generates a soft mask for filtering. In concurrent work, a similar system, based on dilated convolutions and a bidirectional LSTM,

demonstrates good results in unconstrained environments [25]. The model proposed by Gabbay et al. [29] surpasses previous audio-visual speech enhancement results, with a model constructed in encoder-decoder fashion.

For deep learning algorithms it can also be challenging to improve on both intelligibility and quality [30]. In order to overcome this challenge, recent studies include the processing of the phase spectra in order to achieve better speech quality [31], [32], [33]. Alternatively, other studies incorporate perceptual measures into the training method itself [34], [35].

Ideally, deep learning models should be trained using data that is representative of the settings in which they are deployed. This means that in order to achieve a good performance in a wide variety of settings, very large audio-visual datasets for training and testing need to be collected. In practice, to overcome this problem, some systems are trained using a large number of complex acoustic scenes that are synthetically generated by adding target speech signals and signals from sources of interference at several SNRs. This way of generating synthetic training material has empirically shown its effectiveness in both audio-only and audio-visual settings. Speech signals processed this way improve in terms of both estimated speech quality and intelligibility [1], [28], [36]. However, such models comprise of very large number of parameters and require a very high computational power. On the other hand, small networks, which are suitable for more applications, usually succeed to enhance only specific speakers which they were trained on. We say that such models are speaker dependent.

2.2 Speaker dependency

Speech enhancement algorithms tend to suffer from the problem of speaker dependency. Thus, in order to enhance a target speaker, the model must have a previous familiarity with the speakers in the video. A classical speech enhancement model should ideally have prior clean samples of the target speaker only. Moreover, its ability to suppress the noise without introducing perceptible distortion to the speech signal is also affected by assumptions made on the background noise [3], [15].

Deep learning methods also need prior information about the target speaker. Supervised deep learning models learn how to perform speech enhancement after a training procedure, in which they are presented with pairs of degraded and clean speech signals, together with the video of the speakers. Therefore, small neural networks, which comprise of small number of parameters, can only learn a feature embedding from relatively small datasets, and thus suffer from speaker dependency. Other models, such as the models presented by Ephrat et al. [1] and by Afouras et al. [28], avoid speaker dependency by comprising a very large number of parameters, allowing them to learn more generic feature embedding. However, large models require very large database of speakers and their training procedure requires a high computational power, tuning lots of hyperparameters. Eventually, such large complex models cannot be integrated into mobile or embedded devices.

Building a speaker-independent system is also a challenge usually encountered in the task of speech reconstruction from silent videos. Many speech reconstruction systems are personalised

[2], [37], meaning that they are trained and deployed for a particular speaker. Other systems are conditioned on speaker embeddings which are extracted from a reference speech signal [38]. To tackle this problem, a generative adversarial network (GAN) that can directly estimate time-domain speech signals from the video frames of the speaker's mouth is proposed by Vougioukas et al. [39]. Although this approach is capable of reconstructing intelligible speech also in a speaker independent scenario, the speech quality achieved is relatively low and the generated speech signals are characterised by a low-power hum.

This challenging problem of speaker dependency prevents speech enhancement algorithms from running in real-time applications, where a previous familiarity is not guaranteed. Moreover, for most real-time applications of speech enhancement, it is important to derive a solution which is effective and low-power, enabling the deployment in mobile systems, such as hearing aids.

2.3 Few-Shot Learning

Deep neural networks are usually trained using large quantities of labeled data which is tailored to the task. However, in many practical applications, only a handful of examples are available. Moreover, the process of labeling is expensive and time-consuming. Furthermore, once a network is trained, the number of categories remains fixed. Thus, in order to expand the set of recognizable categories, the computationally expensive training procedure needs to be restarted together with a sufficiently large amount of labeled examples of the new category [40].

Unfortunately, neural networks tend to struggle when data is scarce or when they need to adapt to new tasks within a few numbers of steps. On the contrary, humans are able to learn new concepts quickly, given just a few examples. This performance gap between human and artificial learners is usually explained as that humans can effectively utilize prior experiences and generalize their knowledge when learning a new task, whereas artificial learners usually overfit without the necessary prior knowledge.

This highly challenging key problem is called few-shot learning. It is targeting knowledge transfer from sets with abundant data to other sets with only few available examples. In image classification tasks, we denote the former sets as *base* classes (or categories), and the latter sets as *novel* classes. A few-shot learning task is characterized by a number of ways, which is the number of novel classes, and a number of shots, which means the number of novel training samples available of each class. The standard examined cases of few-shot learning in image classification are the case of *5-way 1-shot* and of *5-way 5-shot*.

In a few-shot classification problem, the model tries to optimize a decision boundary for each task with just a few data samples of each class. Research literature on few-shot learning exhibits great diversity, but can be mainly divided into 3 major approaches: data augmentation (example synthesis), metric-learning and meta-learning (learning-to-learn).

Data augmentation (example synthesis) involves with techniques to hallucinate or generate more data samples of the novel classes [41], [42], in order to diminish the imbalance between base and novel classes.

Metric-learning approaches attempt to learn feature representations that preserve the class neighborhood structure, thus features of the same class are closer to each other than features of different classes. It practically creates a feature embedding space of the training data with intra-class similarity and inter-class dissimilarity. The model proposed in [43] is trained with a weighted classifier using an attention mechanism, that is applied to the output of a feature embedding which is trained on the base set. Another notable metric-learning model proposed by Snell et al. [44] is trained with episodic sampling and a loss function based on the performance of a nearest-mean classifier (Euclidean distance to the mean) applied to a few-shot training set. The model introduced in [5] generates classification weights for a novel class based on a feature extractor using the base training set. Siamese neural networks are used by Koch et al. [45] to rank similarity between inputs and essentially learn one metric for all the tasks. More metric-learning methods worth mentioning involve CNN-based relation modules [46] and graph neural networks [47].

Meta-learning (learning-to-learn) has been proposed as a framework to address the challenging few-shot learning setting [48]. Meta-learning tries to learn a shared strategy across different tasks to form decision boundaries from few samples, in the hope that this strategy is able to generalize to new tasks. The key idea is to leverage a training on large number of similar few-shot tasks in order to learn how to adapt a base-learner to a new task for which only a few labeled samples are available. As opposed to transfer learning or domain adaptation [49], [50], where a network is trained in one domain and then transferred (with some fine-tuning) to a different feature space which might also have another distribution, in meta-learning techniques the network is trained on many meta-tasks in order to generalize on tasks.

A meta-training example is a classification task T sampled from a distribution $p(T)$. T is called episode, including a training split T_{train} to optimize the base-learner, and a test split T_{test} to optimize the meta-learner. Meta-training phase aims to learn a meta-learner from a number of episodes $\{T\}$ sampled from $p(T)$. In each episode, meta-training has a two-stage optimization. The first stage is called base-learning, where the loss is used to optimize the parameters of the base-learner. The second stage contains a feed-forward test on episode test data points. The test loss is used to optimize the parameters of the meta-learner. Thus, the number of meta-learner updates equals to the number of episodes. The meta-test phase aims to test the performance of the trained meta-learner for its fast adaptation to new tasks. During this phase, the meta-learner teaches the base-learner to adapt to an unseen task T^{new}. The final evaluation of the meta-learning process is done by testing a set of unseen data points T^{new}_{test} or averaging results of multiple unseen sets $\{T^{new}_{test}\}$.

The model proposed by Finn et al. [51] aims to meta-learn an initial condition (set of neural network weights) that is good for fine-tuning on few-shot problems. The strategy proposed is to search for the weight configuration of a given neural network such that it can be effectively fine-tuned on a sparse data problem within a few gradient-descent update steps. A similar model, introduced in [52], uses a task specific initial condition and performs the adaptation in a lower-dimensional space. The learner model in [53] is adapted to a new episodic task by

a recurrent meta-learner, based on LSTM, producing efficient parameter updates. Thus, the model not only includes meta-learning an optimal initial condition, but also an optimizer that is trained to be specifically effective for fine-tuning. The meta-learner proposed by Santoro et al. [54] is trained to represent entries from a sample set in an external memory, offering the ability to quickly encode and retrieve new information, and hence to rapidly assimilate new data and leverage this data to make accurate predictions after only a few samples.

Few-shot learning, while a well-studied problem, is still a challenge for neural networks, and there is a long way to go before demonstrating abilities similar to humans [55]. In this research we exploit meta-learning approaches for avoiding speaker dependency in the task of audio-visual speech enhancement. Alongside its impressive advantages, we examine the drawbacks of the proposed method in the context of speech enhancement. Furthermore, we investigate limitations of a few-shot learning model in its original task of image classification. We then propose a solution for the drawbacks.

Chapter 3

Fast Adaptation Network for Few-Shot Audio-Visual Speech Enhancement

3.1 Introduction

Speech is one of the primary ways to communicate, but audio signals are often recorded in noisy environments. People are able to focus their attention on a particular speaker, mentally "muting" all other speakers and background noises. This capability comes natural to us humans. Known as the cocktail party effect [56], automatic speech separation, although a well-studied problem, is still a challenge for computers. Audio-visual speech enhancement is the problem in which given a noisy audio signal and the corresponding speaker video frames, the model would produce an enhanced audio signal containing only the target speaker's voice, whereas the rest of the speakers and background noise would be suppressed.

Traditionally, speech enhancement has been tackled using signal processing and machine learning techniques applied to the speech signals solely. Audio based methods isolate multi-talker simultaneous speech, using voice characteristics of a known speaker [57, 58, 59, 60]. Deep learning audio-only methods are widely used for speech denoising as well as for speech separation tasks. Wang and Chen [17] give a comprehensive overview of those methods.

Recent works use visual information from the target speakers to improve speech enhancement results. Intuitively, movements of a person's mouth, for example, should correlate with the sounds produced as that person is speaking, which in turn can help identify which parts of the audio correspond to that person. Ephrat et al. [1] show that adding the visual signal as input not only improves the speech separation quality significantly in cases of mixed speech, but also associates the separated, clean speech tracks, with the visible speakers in the video.

Recently, there is an increased interest in using neural networks with audio-visual inputs for speech separation and enhancement [1, 25, 28]. The idea is based on the fact that the visual aspect of speech, which is essentially unaffected by the acoustic environment, can be efficiently

fused by exploiting the flexibility of data-driven approaches, specifically deep learning. The method proposed in [29] surpasses previous audio-visual speech enhancement results, with a model constructed in encoder-decoder fashion. Each modality of the input enters a separate encoder which consists of a dual tower convolutional neural network. Then, both outputs are combined as a shared embedding representing the audio-visual features. The shared embedding enters a deconvolutional decoder which produces a spectrogram representing the enhanced speech.

Although demonstrating impressive results, the model of [29], together with other speech enhancement models, suffer from the problem of speaker dependency. Thus, in order to enhance a target speaker, the model must have previous familiarity with the speakers in the video. Ideally, its training set should comprise of samples of the target speaker only. This challenging problem of speaker dependency prevents speech enhancement algorithms from running in real-time applications, such as hearing aids, where a previous familiarity is not guaranteed.

In this work, we introduce a fast adaptation speech enhancement (FASE) network for overcoming speaker dependency. The proposed model is inspired by approaches to the problem of few-shot learning in image classification tasks. Few-shot learning [4] is a task in which a classifier must adapt quickly to accommodate new classes from only a few examples. The class that is underrepresented in the training set is called *novel* category, and the other ones are denoted as *base* categories. The goal is to efficiently learn novel categories from only a few training samples, and nonetheless keep high performance on the base categories which the model was trained on. Few-shot learning is challenging, since the network must integrate its prior knowledge with a small amount of new information, while avoiding overfitting due to the large number of parameters compared to the relatively small amount of new data. Therefore, a naive approach, such as retraining the model on the new data, might severely overfit the new data. Gidaris et al. [5] show a trade-off between the performance on novel and base categories in image classification. As the number of training samples per novel category increases, the performance of a few-shot learning network on novel categories improves, whereas there is a performance degradation on the base categories.

A notable approach to solving few-shot learning tasks is meta-learning. Meta-learning [61] has emerged as a promising approach for enabling systems to quickly learn new tasks by building upon experience from previous related tasks [62, 45, 54]. To accomplish this, meta-learning methods explicitly optimize for few-shot generalization across a set of meta-training tasks. The meta-learner is trained such that the trained model can quickly adapt to new tasks using only a small number of examples or trials [53, 51]. Actually, the meta-learning problem treats an entire set of few-shot learning tasks as training set of the meta-learner.

Our FASE model combines methods of meta-learning with the task of audio-visual speech enhancement, in order to overcome the problem of speaker dependency. It comprises an inner network, based on a deep encoder–decoder architecture, which is trained by an outer neural network model. This outer meta-learner is trained to learn how to update the parameters of the inner model for fast adaptation to new speakers. Unlike speaker dependent models, where in each training phase the input data must include videos of the target speaker solely, in our

model the training set includes a variety of speakers with highly imbalanced number of samples.

The proposed FASE model outperforms prior art on TCD-TIMIT [63] dataset, designed for the task of audio-visual speech recognition and lip reading. The improvements in PESQ and STOI measurements apply both for the novel speakers (those with only few training samples) and for the base speakers (with a large representation in the training set). We also demonstrate improvements in inference compute time. To our knowledge, FASE model is the first meta-learning architecture to overcome the problem of speaker dependency in audio-visual speech enhancement, using meta-learning approaches.

The rest of this chapter is organized as follows. In Section 3.2, we demonstrate the problem of speaker dependency in prior art. In Section 3.3, we introduce the FASE model, describe its architecture and methodologies. Section 3.4 presents experimental results of our model. We further investigate cases with larger numbers of novel samples in the training set. Conclusions are presented in Section 3.5.

3.2 Speaker Dependency

The neural network model for audio-visual speech enhancement proposed by Gabbay et al. [29] outperformed previous art. The authors compared their model to Vid2speech [64] and Gabbay [27], using an objective perceptual evaluation of speech quality (PESQ) [65]. They trained the networks on two audio-visual datasets, GRID [66] and TCD-TIMIT [63]. In all cases, background speech was sampled from the LibriSpeech [67] dataset. In each experiment the training set, and accordingly the test set, included video samples of the target speaker only. These experiments hint that the model is speaker dependent.

In order to further investigate the results in [29], and to show that the model is indeed speaker dependent, we perform extensive experiments of this model on several different settings. All experiments are conducted according to the parameters and model described by Gabbay et al. [29]. In this section, we present our setting and analyze the results.

We use TCD-TIMIT lip speakers' dataset [63] both for random target speakers and for background speech noise. This background noise is more challenging than LibriSpeech [67] dataset, because of the similarity in characteristics to the target speaker's speech signal. As described by Gabbay et al., we divide the data modalities (audio samples and video frames) into train set, validation set, and test set by 70%, 20%, 10%, respectively. Every time we train the network model end-to-end, as in [29].

3.2.1 Experimental Settings

We conduct two sets of experiments. The first set of experiments (I, II, III) checks the setup of one single target speaker in the input of all learning phases (training, validation and test). The second set of experiments (IV, V, VI) was conducted in order to assess the performance when two speakers are shuffled in the input of all phases, describing a more challenging environment for speaker dependent networks.

Experiment I – This is the classical setup of a speaker dependent network. We take all data of one speaker as the target speaker and another speaker as the background noise. Thus, both the target speaker and the noise are presented during train, validation and test.

Experiment II – We take data of one speaker as the target speaker and another speaker as the background noise. Then, during test phase we use the same target speaker, but now with different background noise (of a third speaker). The purpose of this experiment is to check whether the network is dependent not only on the target speaker, but also on the background noise.

Experiment III – For training, we use data of one speaker as the target speaker and another speaker as the background noise. For test set we use dataset of a third speaker as target speaker. The noise of all phases is taken from the same background speaker. Thus, the target speaker changes at test phase but the noise source does not change. We check whether the network has the opportunity to learn how to separate a specific noise from an unknown speaker.

We further check the speaker dependency of the model with two target speakers as input data. The input samples of the speakers is randomly shuffled, in both training, validation, and test sets.

Experiment IV – The same two target speakers are in the training, validation and test sets. We use the same background noise of a third speaker in the training, validation, and test sets. Thus, both the target speakers and the noise are presented during all learning phases. This experiment resembles Experiment I, for validating the speaker dependency of the model in the classical conditions, but now with two shuffled input speakers.

Experiment V – Similarly to Experiment II, we check whether the background noise is learned specifically. We shuffle two speakers as input for the training set, both with the same dataset for background noise. For test set we take the same speakers but now with a different speaker as the background noise.

Experiment VI – In order to check the dependency on the target speaker regardless of the noise, we conduct an experiment that resembles Experiment III. For training set we randomly choose two shuffled speakers with the same speaker as background noise. Then during test another speaker is chosen as the target speaker, but the noise does not change.

3.2.2 Dependency Results

We show a comparison of results of the single speaker experiments (I,II,III) in Table 3.1. We observe that the PESQ results are aligned with the audio quality obtained in the enhanced output. The first experiment validates the results presented in [29], where mean PESQ of 2.85 is reported. Looking at Table 3.1, it can be seen that there is a large dependency between the speaker in the training set and the speaker in the test set. When they are sampled from dataset of the same speaker, such as in Experiment I, there is a significant improvement, as opposed to Experiment III. Furthermore, it is noticeable that the background noise is somewhat also learned during training phase. When the noise of the test set is sampled from dataset of another speaker, the PESQ results decrease, as in Experiment II compared to Experiment I.

Table 3.1: Results of Single Speaker Experiments.

	Mean PESQ	Mean STOI
Experiment I	2.86	0.254
Experiment II	2.105	0.249
Experiment III	1.989	0.3

Table 3.2: Results of Two Speakers Experiments.

	Mean PESQ	Mean STOI
Experiment IV	2.291	0.223
Experiment V	1.981	0.326
Experiment VI	1.552	0.28

Table 3.2 summarizes the results of two-speakers experiments (IV,V,VI). It can be seen that the conclusions from the single speaker experiments are also relevant in the case of more than one target speaker. The results validate the great dependency of the model [29], both on the target speaker and on the background noise.

We conclude that in both sets of experiments, PESQ and STOI objective measurements demonstrate the dependency of the model not only on the target speaker but also on the background noise, validating that the model is indeed speaker dependent. This dependency is a great disadvantage of [29], which prevents it from performing in real-life situations.

Our FASE model tackles the problem of speaker dependency, proposing a different approach to learn the task of speech enhancement. FASE model avoids speaker dependency and enables its use in real-time applications.

3.3 Methodology

FASE model optimizes the audio-visual speech enhancement model for fast adaptability over episodes, allowing it to quickly adapt to new tasks with only a few examples. The setting of our task includes an outer network which learns how to update an inner audio-visual speech enhancement network, to succeed with a small number of training samples. The whole model was trained end-to-end, but for convenience of the readers we divide the explanation of our training methodology into several sections.

In Section 3.3.1, we start by describing the setup of the few-shot learning task, which allows checking results of novel and base speakers. We continue with explanation about the pre-processing of the data. Section 3.3.2 concentrates on the meta-learning methodology, which allows learning to overcome the problem of few-shot speech enhancement. In Section 3.3.3, we describe in details two proposed architectures for the inner network, compare them and describe the data flow in each of them.

3.3.1 Task Setup

The examined task includes data of several speakers, where one of the speakers has only few video samples in the training set. Similar to few-shot learning in image classification, we denote the this speaker as *novel* and the other speakers as *base*. The goal of our FASE model is to succeed enhancing the novel speaker in the test set, without performance degradation on the base speakers.

We developed an automatic task-creation method for episode learning. It chooses a random novel speaker, divides data into training, validation, and test sets, by 70%, 10%, and 20% accordingly. Then, for each one of the sets of this task, it picks a new random speaker and inserts its audio samples as background noise. We chose the noise to be speakers and not some arbitrary noise, for achieving a much more challenging task which is also very common in real-life situations.

For each noisy video, the input of the neural network contains both the video frames and a spectrogram of the noisy audio. The output is a spectrogram of the enhanced speech. During pre-processing stage, the audio signal is re-sampled to 16 kHz and short-time Fourier transform (STFT) is applied to the waveform signal. The window size of the applied STFT corresponds to the length of a single video frame. The phase is kept aside for reconstruction of the enhanced signal spectrogram. The magnitude is multiplied by a mel-spaced filterbank to form a log mel-scale spectrogram, which serves as the input to the audio encoder. The spectrogram is sliced to pieces which correspond to a batch of consecutive images.

Pre-processing of the video includes re-sampling it to 25 fps and dividing to non-overlapping segments of frames. Input images are cropped around the mouth of the speaker, in order to capture visual features from the target speaker's lips. The cropped images allow the network to recognize speech visually without overfitting on specific speakers. The images are normalized and divided into batches that correspond to the audio input segments.

Both the audio and video encoders, as part of the inner network, are trained end-to-end by the outer network.

3.3.2 Meta-Training

Our training methodology consists of episodes of tasks. Each episode is a setup in which a random speaker is the novel category, and the inner network runs for 200 epochs (or until convergence on the validation set). The episodes allow the outer network to learn how to tackle the few-shot learning task. It can be interpreted as the outer network's data are the few-shot learning tasks, and its goal is generalizing for success on a similar new task.

The complete FASE model training is described in Algorithm 3.1. At the beginning, the number of novel samples, K, is chosen similarly to k-shot learning tasks. Then, the model randomly constructs M meta tasks. Each task contains a training set T, with base speakers and only K samples of a randomly picked speaker, and corresponds to a validation set V. The validation set includes an equal number of novel and base samples, in order to imitate the test set and examine all speakers evenly. The model is parameterized by some parameter vector θ,

Algorithm 3.1 FASE Model

Require K: Number of novel speaker training samples
Randomly construct M meta tasks, each comprises of a training set T and a validation set V
Init model weights θ
Init model hyperparameters α_{in}, α_{out}
while not done **do**
  Sample batch of meta-tasks $M_i \in M$
  for each task M_i **do**
    Sample batch of training data $T_i \in T$ and a batch of validation data $V_i \in V$
    for each pair $T_i \in T$, $V_i \in V$ **do**
      Forward pass T_i in inner network (audio, video)
      Compute training loss $\mathcal{L}_{T_i}(f_\theta)$
      Evaluate $\nabla_\theta \mathcal{L}_{T_i}(f_\theta)$
      Update $\hat{\theta}_i \leftarrow \theta - \alpha_{in} \nabla_\theta \mathcal{L}_{T_i}(f_\theta)$
      Compute validation loss $\mathcal{L}_{V_i}(f_\theta)$
      Update lr
    end for
    Compute task loss $\mathcal{L}_{M_i}(f_{\hat{\theta}_i})$
  end for
  Evaluate task gradient $\nabla_\theta \sum_{M_i \in M} \mathcal{L}_{M_i}(f_{\hat{\theta}_i})$
  Update $\theta \leftarrow \theta - \alpha_{out} \nabla_\theta \sum_{M_i \in M} \mathcal{L}_{M_i}(f_{\hat{\theta}_i})$
end while

assuming that the loss function f_θ is smooth enough in θ for gradient-based learning techniques. When adapting to a new task M_i, the model's parameters are updated. In practice, training is done with Adam optimizer. We define separated step sizes, α_{in} and α_{out}, for the inner and outer networks. The network is trained end-to-end with mean-squared error (MSE) loss, calculated between the clean target speech spectrogram and the output spectrogram.

3.3.3 Network Configurations

We propose two architectures of the inner network. First inner network resembles the architecture of Gabbay et al. [29], and the model which includes this inner network is denoted as FASE-baseline. Second inner network configuration is a less complex, much faster network. Therefore we denote the model with the second architecture as FASE-opt. Both configurations are based on encoder-decoder model, where the encoder consists of a dual tower convolutional neural network, and the decoder is a mirror of the audio encoder.

First, we describe the baseline configuration. The audio and video inputs enter two separate encoders. The audio encoder consists of 5 convolution layers, each followed by batch normalization and leaky-ReLU non-linearity. The video encoder consists of 6 convolutional layers, each followed by batch normalization, leaky-ReLU, max pooling, and dropout. In order to represent the audio-visual features of the video the outputs of the two encoders are concatenated into a shared embedding . The decoder module consists of 5 transposed convolutional layers followed by batch normalization and leaky-ReLU, thus mirroring the audio encoder. This module decodes

the shared embedding into a spectrogram representing the enhanced speech. Passing 3 consecutive fully-connected layers, the resulting vector is then ready to be decoded. In order to attain the output enhanced video, the last layer is of the same size as the input spectrogram, which is then constructed back together with the images. This baseline configuration comprises 22M trainable parameters, among them 13M in the dual encoder and 8M in the decoder module.

FASE-opt inner network configuration is based on the baseline configuration, but is much shallower and computationally simpler. The goal of developing the opt configuration was to avoid overfitting that might appear due to the small number of training samples and the imbalance between categories. We examined many overfitting solutions and arrived to the final opt configuration. Its audio encoder consists of 3 convolutional layers, with regularization on kernel and bias, and a layer of max pooling between any two of them. The video encoder also has regularization of kernel and bias in each of its 6 layers, followed by batch normalization, leaky-ReLu non-linearity, max pooling and a spatial dropout. The concatenation and fully-connected layers remain untouched for receiving the shared audio-visual embedding. The audio decoder corresponds to the audio encoder and thus comprises up-sampling layers between any 2 of its 3 deconvolutional layers. Each transposed convolution is followed by batch normalization and leaky ReLu non-linearity. All in all, the opt configuration comprises of 12M trainable parameters, among them 9M in the encoder module and 3M in the decoder. This simpler configuration enables faster adaptation and convergence without overfitting, as presented in the next section.

3.4 Results

In this section, we demonstrate the proposed model's performance. First, we check whether the outer network indeed trains the inner network to succeed in the few-shot speech enhancement tasks. Then, we compare three audio-visual speech enhancement models: The model of Gabbay et al. [29], which surpassed previous ones in the case of one speaker, and our proposed models, FASE-baseline, and FASE-opt.

3.4.1 Dataset

The TCD-TIMIT dataset [63] was constructed for audio, visual, and joint audio-visual experiments. It consists of three lip speakers, specially trained to speak in a way that helps the deaf understand their visual speech, as well as 59 volunteer speakers with around 200 videos. The dataset contains videos of 2 seconds long recorded from two angles: videos in which the speaker is facing the camera, and videos with face direction of 30 degrees. We chose to use the front facing videos, assuming the speech features are more significant in them. All speakers are recorded saying sentences from the TIMIT dataset, reaching a total of 6913 phonetically rich sentences.

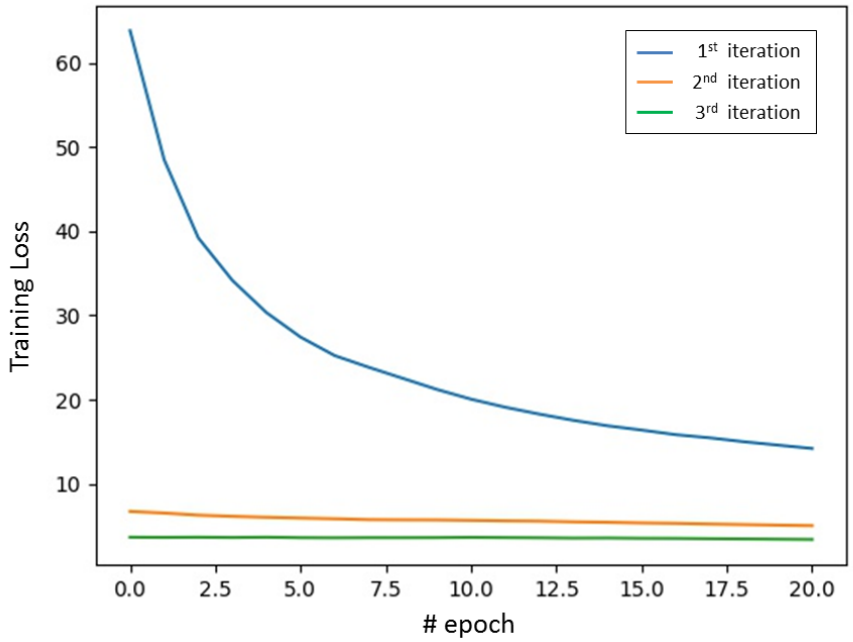


Figure 3.1: Validation of convergence during outer network iterations. First iteration is in blue, second one in orange and third one in green. Each iteration starts with a smaller loss than the previous one, and the loss decreases along the inner network training.

3.4.2 Meta-Training

We validate that FASE models meta-training trains the inner network as expected. Figure 3.1 shows the training loss in 3 consecutive meta-iterations. It can be seen that first iteration starts with largest loss value and decreases during training of the inner network on the first task. The second iteration of the outer network, which starts after updating the network parameters, begins with a smaller loss value and decreases further. The same phenomenon applies to the next iteration. Therefore, we conclude that the outer network indeed learns the update rule of the inner network in order to succeed in the different few-shot learning tasks.

3.4.3 Few-Shot Speech Enhancement Results

We continue examining FASE models performance on the test set in the setup of few-shot learning. We start by checking the cases of 1-shot, 5-shot, and 10-shot learning. We then continue investigating the performance with larger values of K, which are not standard cases for few-shot learning tasks. We expect the speaker dependent model of Gabbay et al. [29] to succeed better in larger values of K.

All results refer to a random test set with equal number of base speakers and novel speakers samples. Few-shot learning models usual suffer from a trade-off between the performance on the novel categories and the base categories. Therefore, we validate both the ability to dynamically learn novel speakers, and the capability not to forget the base ones.

Table 3.3: Few-Shot Learning Results.

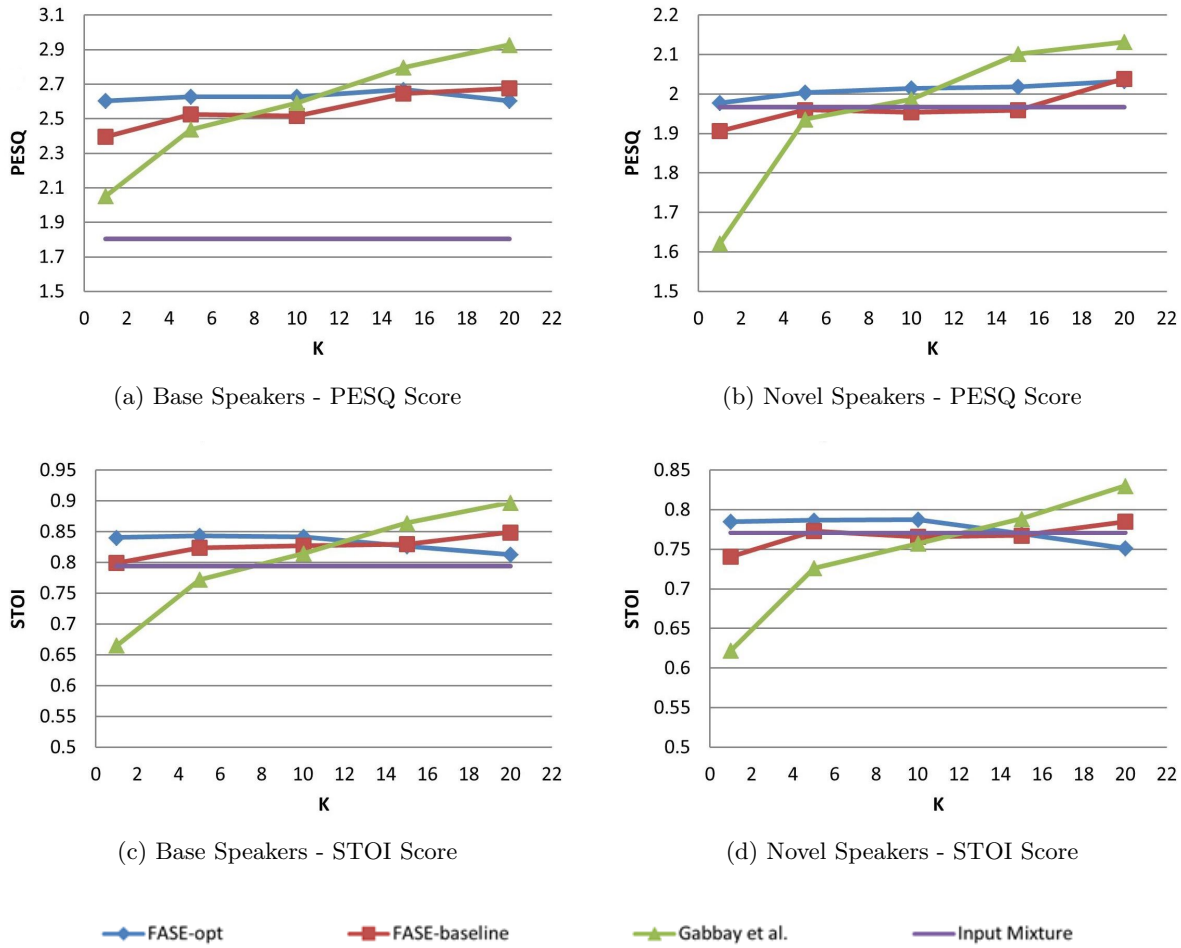
K	Model	Base Speakers		Novel Speakers	
		PESQ	STOI	PESQ	STOI
1	Gabbay et al. [29]	2.052	0.665	1.621	0.622
	FASE-baseline	2.397	0.799	1.906	0.741
	FASE-opt	**2.603**	**0.840**	**1.977**	**0.785**
5	Gabbay et al. [29]	2.438	0.772	1.936	0.726
	FASE-baseline	2.526	0.824	1.960	0.773
	FASE-opt	**2.628**	**0.844**	**2.00**	**0.787**
10	Gabbay et al. [29]	2.590	0.815	1.986	0.757
	FASE-baseline	2.516	0.827	1.953	0.766
	FASE-opt	**2.628**	**0.842**	**2.014**	**0.787**

Table 3.4: Results With a Larger Training Set.

K	Model	Base Speakers		Novel Speakers	
		PESQ	STOI	PESQ	STOI
15	Gabbay et al. [29]	**2.797**	**0.864**	**2.101**	**0.788**
	FASE-baseline	2.647	0.830	1.959	0.767
	FASE-opt	2.668	0.826	2.018	0.769
20	Gabbay et al. [29]	**2.928**	**0.897**	**2.132**	**0.830**
	FASE-baseline	2.677	0.849	2.039	0.785
	FASE-opt	2.604	0.813	2.033	0.752

The enhanced speech of both base and novel speakers was evaluated by two objective measures, perceptual evaluation of speech quality (PESQ) [65] and short-time objective intelligibility measure (STOI) [68]. The PESQ measure was found to correlate well with a subjective listening test using Multi-stimulus test, conducted with Hidden Reference and Anchors (MUSHRA), and to be able to well predict separation quality. The STOI measure was found to be well suited to predict the speech intelligibility result. Therefore, these measures were recommended to evaluate single-channel speech separation algorithms [69].

Table 3.3 describes PESQ and STOI results of each of the networks in the cases of $K = 1, 5, 10$. Figure 3.2 shows all the measurements of all models with different values of K. In the cases of 10-shot learning or less, FASE-opt algorithm outperforms others in all experiments and measurements. The small number of parameters allows it to converge fast and accurately for small numbers of K. The trade-off between performance on novel speakers and base speakers is imperceptible, as FASE models outperform with both novel and base speakers. The performance gap between the three models is most significant in the case of 1-shot learning, emphasizing the ability of FASE-opt model to run online when the model is acquainted with a new target speaker. We conclude that FASE-opt is a state-of-the-art speech enhancement algorithm when only few samples of the target speaker are available for training.



(a) Base Speakers - PESQ Score

(b) Novel Speakers - PESQ Score

(c) Base Speakers - STOI Score

(d) Novel Speakers - STOI Score

Figure 3.2: Summary of PESQ and STOI results of base and novel speakers with different number of training samples of the novel speaker, $K = 1, 5, 10, 15, 20$. Results of Gabbay et al. are in green, FASE-baseline in red, FASE-opt in blue, and input mixture in purple. FASE-opt surpasses others in all measurements of the few-shot learning cases, but Gabbay et al. improves when training set of the novel speaker enlarges to non few-shot learning cases.

3.4.4 Larger Training Set Results

We further investigate the models with larger values of K, which are not standard for few-shot learning algorithms. As we assumed, the advantages of meta-learning networks are not realized in those cases, whereas the speaker dependent network improves as the training set of the target speaker increases. Table 3.4 describes the performance of the three models with $K = 10$ and $K = 15$, and Fig. 3.2 shows results with all values of K. It can be seen that in the cases of $K = 10$ and $K = 15$, the model in [29] outperforms FASE models. The diminished imbalance between speakers, contributes to the performance of the speaker dependent network both on novel speakers and on the base speakers. We also note that FASE-baseline succeeds better than FASE-opt in those cases, indicating that the larger number of parameters enables it to better capture features of this larger training set. We conclude that FASE models are more suitable

for their original purpose, tackling the problem of too little previous familiarity with the target speaker.

3.5 Conclusions

We have addressed the real-life problem of audio-visual speech enhancement when there is not enough previous familiarity with the target speaker. We introduced FASE model, inspired by meta-learning methods, used originally for the task of few-shot image classification. The model comprises a deep encoder-decoder inner network, which is trained by a meta-learner for fast adaptation to new speakers. Our FASE model outperforms prior art in objective measurements in the cases of few-shot learning, both on novel speakers and on base speakers. Moreover, the small number of training parameters allows FASE model to converge quickly, and be compatible for a variety of mobile health and media applications.

There are many possible applications for the FASE model. One application is pre-processing for automatic speech recognition (ASR). Even though machines can recognize speech relatively well in noiseless environments, there is a significant deterioration in performance for recognition in noisy environments [70]. The enhancement model could address this problem, and improve, for example, ASR for mobile phones in a crowded environment. Moreover, handling overlapping speakers is a known challenge for automatic captioning systems, and separating the audio to distinct sources could help in presenting more accurate and easy-to-read captions. Another important application of such mobile speech enhancement model, is hearing aids. More potential applications are in personalized advertisement, virtual assistants, and applications for intelligence corps, health and wellness.

Chapter 4

From Few Shots to More Shots: An Algorithm to Overcome Few-Shot Methods Limitations

4.1 Introduction

In Chapter 3 we introduced a new neural network model to avoid speaker dependency in audio-visual speech enhancement tasks. The proposed fast adaptation speech enhancement (FASE) model is inspired by methods which were originally developed for the task of few-shot learning in image classification. The examined speech enhancement tasks include data of several speakers, where one of the speakers has only few video samples in the training set. Similar to few-shot learning in image classification, we denote the this speaker as *novel* and the other speakers as *base*. In Section 3.4 we analyzed the results of the model on base and novel speakers, both in the standard few-shot learning case, of few training samples of the novel speaker, and in the case of a larger training set of the novel speaker. Even though FASE model outperforms previous art in the standard few-shot cases, which emphasizes its potential use in real-time applications, it suffers from a performance degradation as the number of novel training samples (shots) increases.

Few-shot learning methods are designed and customized especially for the cases of few novel training samples. In image classification tasks, this implies the cases of 1-shot and 5-shot solely. However, in reality, there is no guarantee for this small number of shots. Ideally, a few-shot learning model should succeed in all cases, with no dependency in the number of training samples of the novel categories. Intuitively, like other neural networks, we would expect to see that as the number of shots increases, the performance of the network on novel categories improves, without harming the performance on base categories. However, achieving successful results with more shots seems to be a significant challenge for few-shot learning methods.

In this chapter we extend the investigation of few-shot learning models, coping not only with very few shots. We choose to concentrate in a successful few-shot learning model for the task of image classification [5], and examine its performance in the cases of larger number of shots.

After discussing and better understanding the disadvantages of few-shot learning methods, we propose a solution. Our approach is to allow better absorption of the few novel category samples in the feature space of base samples. We present a proof-of-concept which manages to achieve a stability and a decreased dependency of the performance on the number of novel training samples. The proposed method improves the accuracy results for the standard few-shot cases. Moreover, our solution is general and can be adapted to other few-shot learning tasks.

The rest of this chapter is organized as follows. In Section 4.2 we provide an overview of few-shot learning methods for image classification. Section 4.3 describes in details the methodology proposed by Gidaris et al. [5] and investigates its performance limitations as the number of shots increases. In Section 4.4 we propose a solution for the limitations and examine it in the standard cases of few-shot learning compared to cases of more shots.

4.2 Related Methodologies

Few-shot learning methods which do not include generating more data, can be divided into two main groups of approaches, which have recently made significant progress. We present both approaches and then discuss common practices in few-shot learning research.

4.2.1 Metric-learning based approaches

Metric-learning approaches attempt to learn feature representations that preserve the class neighborhood structure, such that features of the same object are closer than features of different objects. One notable work is Matching Networks [43], which uses an attention mechanism over a learned embedding of the labeled set of examples to predict classes for the unlabeled points. Matching Networks can be interpreted as a weighted nearest-neighbor classifier applied within an embedding space. The authors of [45] train a Siamese Neural Networks to identify input pairs according to the probability they belong to the same class, thus learning to compute the similarity between a test example and a training example of a novel category. Another recent work [44], proposed a Prototypical Network that learns a non-linear mapping of the input into an embedding space and takes the prototype of each class as the mean of its support set. Classification is then performed for each embedded query point of a test image by finding the nearest class.

4.2.2 Meta-learning based approaches

Meta-Learning approaches train a meta-learner to learn how to update the parameters of the learner's model [71]. These approaches have been applied for learning dynamically changing recurrent networks, as well as for learning to optimize deep networks [72]. One recent method for few-shot image recognition, proposed training an LSTM [71] based meta-learner that is trained, given as input an episode of a few training examples of a new classification task, to sequentially produce parameter updates to a classifier that will optimize the performance such that it will generalize well to a test-set. This method allows learning both the optimizer and

the weight initialization [53]. Another recent paper [51], simplified the above model by a model-agnostic method that only learns the initial learner parameters. Rather than a learned update, only a few gradient descent steps with respect to those initial parameters are needed for a good performance on the new task. Thus this model does not introduce additional parameters for meta-learning, nor does it require a particular learner architecture.

4.2.3 Common practices in few-shot learning

Notably, several meta-learning as well as metric-learning models utilize sampled mini-batches called episodes during training [43], [51] . Each episode is designed to mimic the few-shot task by sub-sampling classes and data points. The use of episodes makes the training problem more faithful to the test environment, based on the principle "train and test condition must match", and thereby improves generalization.

Another common practice is using an attention based model, for allowing the network to "look" at the memory, which stores useful information for solving the task. The work of [5] involves a meta-learner together and a metric-learning mechanism together with an attention kernel for "looking" at the weights of the learned category training samples.

4.3 Performance Limitations

The few-shot learning model presented by Gidaris et al. in [5] combines the ideas of cosine-similarity recognition and classification weight generator, thus adopting meta-learning together with metric-learning approaches. In this section, we introduce the methodology and provide experimental results to extend the examination of the algorithm in the cases of more novel training data.

4.3.1 Cosine-Similarity Based Recognition Model

In classification neural networks, the standard setting is to estimate the classification probability vector $p = C(z|W^*)$, where z is an extracted feature vector, by computing the raw classification score s_k of each category $k \in [1, K^*]$, using the dot-product operator: $s_k = z^T w_k^*$.

In a few-shot learning model, unifying the recognition of the base and novel categories is essential, in a way that the convolutional network would be able to simultaneously handle the classification weight vectors of both categories. However, according to the small number of examples of novel categories, the unification is not feasible with the dot-product based classifier, which is typically the last linear layer of a classification neural network.

In order to overcome this problem, a technical novelty is presented, using meta-learning approach. The classifier $c(\cdot|W^*)$ is implemented as a cosine similarity function between the feature representations and the classification weight vectors:

$$s_k = \tau \cos(z, w_k^*) = \tau \frac{z}{||z||} \frac{w_k^*}{||w_k^*||} \tag{4.1}$$

where τ is a learnable scalar.

The cosine-similarity based classifier, apart from unifying the recognition of both base and novel categories, also leads to feature representations that are able to better generalize on "unseen" categories [73]. Thus, features learned with the cosine-similarity based classifier generalize on novel categories better than those learned with a dot-product based classifier.

Note that even though cosine-similarity function is well established for classifying a test feature, by comparing it with the available training features vectors [43], here it is used for a different kind of purpose, i.e., replacing the dot-product operation of the last linear layer [74].

4.3.2 Few-shot classification weight generator

The few-shot classification weight generator $G(\cdot,\cdot|\phi)$ gets as input the feature vectors

$$Z^{'} = \{z_i^{'}\}_{i=1}^{N^{'}} \tag{4.2}$$

of the $N^{'}$ training examples. One choice [44] is to infer the classification weight vector $W^{'}$ by averaging the feature vectors of the training examples, in which case

$$W_{avg}^{'} = \frac{1}{N^{'}} \sum_{i=1}^{N^{'}} \frac{z_i^{'}}{||z_i^{'}||} \tag{4.3}$$

and the final classification weight vector is

$$W^{'} = \phi_{avg} \odot W_{avg}^{'} \tag{4.4}$$

where $\odot$ is the Hadamard Product and $\phi_{avg} \in R^d$ is a learnable weight vector.

However, for one-shot learning the averaging might not infer an accurate classification weight vector. Moreover, this approach does not fully exploit the knowledge that the convolutional network acquires during its training. Therefore, Gidaris et al. [5] propose an attention-based weight inference which will exploit the meta-learning approach in order to enhance the feature averaging mechanism. An attention kernel allows "looking" back at the memory that contains the base classification weight vectors. The attention-based classification weight vector is computed by

$$W_{att}^{'} = \frac{1}{N^{'}} \sum_{i=1}^{N^{'}} \sum_{b=1}^{K_{base}} Att\Big(\phi_q \frac{z}{||z||}, k_b\Big) \frac{W_b}{||W_b||} \tag{4.5}$$

where the attention kernel $Att(\cdot,\cdot)$ is implemented as a cosine similarity function, $\phi_q \in R^{d \times d}$ is a learnable weight matrix that transforms the feature vector to query vector used for querying the memory, and $\{k_b\}_{b=1}^{K_{base}} \in R^d$ is a set of K_{base} learnable keys (one per base category) used for indexing the memory. Thus, the final classification weight vector is computed as:

$$W^{'} = \phi_{avg} \odot W_{avg}^{'} + \phi_{att} \odot W_{att}^{'} \tag{4.6}$$

Therefore, the proposed few-shot weight generator exploits the acquired knowledge about the

visual world, represented by the base classification weight vectors, in order to improve the recognition of novel categories.

4.3.3 Experimental Settings

We apply the proposed image classification model of Gidaris et al. [5] in order to validate its ability to dynamically learn novel categories without forgetting the base categories, and further investigate its limitations with more training samples of the novel category. All the experiments are done on the Mini-ImageNet dataset [43], which is a small version of ImageNet dataset for image classification [75]. Mini-ImageNet includes 100 categories with 600 images per category, each of size 84x84. We used the splits proposed in [53], which include 64 categories for training, 16 categories for validation, and 20 categories for testing. As recommended, we randomly divide the samples into training, validation and test sets by 70%, 10%, and 20% accordingly.

As described in [5], the feature extractor used in all experiments is a convolutional network model with 4 convolutional modules, using 3x3 convolutions, followed by batch normalization and 2x2 max-pooling. The first two convolutional layers have 64 feature channels and the latter two have 128 feature channels. According to [5], since the model comprises a cosine classifier, the ReLu non-linearity function is used along the network apart from the last layer.

Few-shot learning settings are denoted by *n-way k-shot*, which means there are k novel categories and n training samples of each. According to the results reported in [5], their configuration achieved state-of-the-art accuracy, which surpass the prior models Matching Networks [43] and Prototypical Networks [44]. However, the paper only describes the standard few-shot settings of *5-way 1-shot*, and *5-way 5-shot* learning. We are interested in checking the limitations and trade-offs of [5] not only in the standard few-shot cases but also with more shots. Therefore we conduct experiments with *5-way k-shot* setting, where k values go as high as 200. Note that the case where $k = 200$ is more shots but there is still an imbalance between the novel categories to the base categories, which comprise of 400 training images each. For all experiments we used the same values of k for both training and test phases.

4.3.4 Experimental Results

The results of our experiments on *5-way k-shot* learning appear in Fig. 4.1. We explore both cases of very few shots and more shots. For standard cases of few-shot learning, i.e. small values of k, the accuracy on base categories is higher than novel ones, although both improve as k increases. In the case of higher values of k, the accuracy on novel categories surpasses the accuracy on base samples, which does not improve. We conclude that there is a dependency on the value of k, which is expressed by the performance trade-off between base and novel categories in the cases of smaller compared to higher values of k. This phenomenon may be caused by the increased number of vectors representing the novel categories, which inclines as k increases. As the network was built for the standard cases of very few shots, the changed imbalance affects the results of the network.

This is a significant limitation of a few-shot learning model, since it only fits limited cases

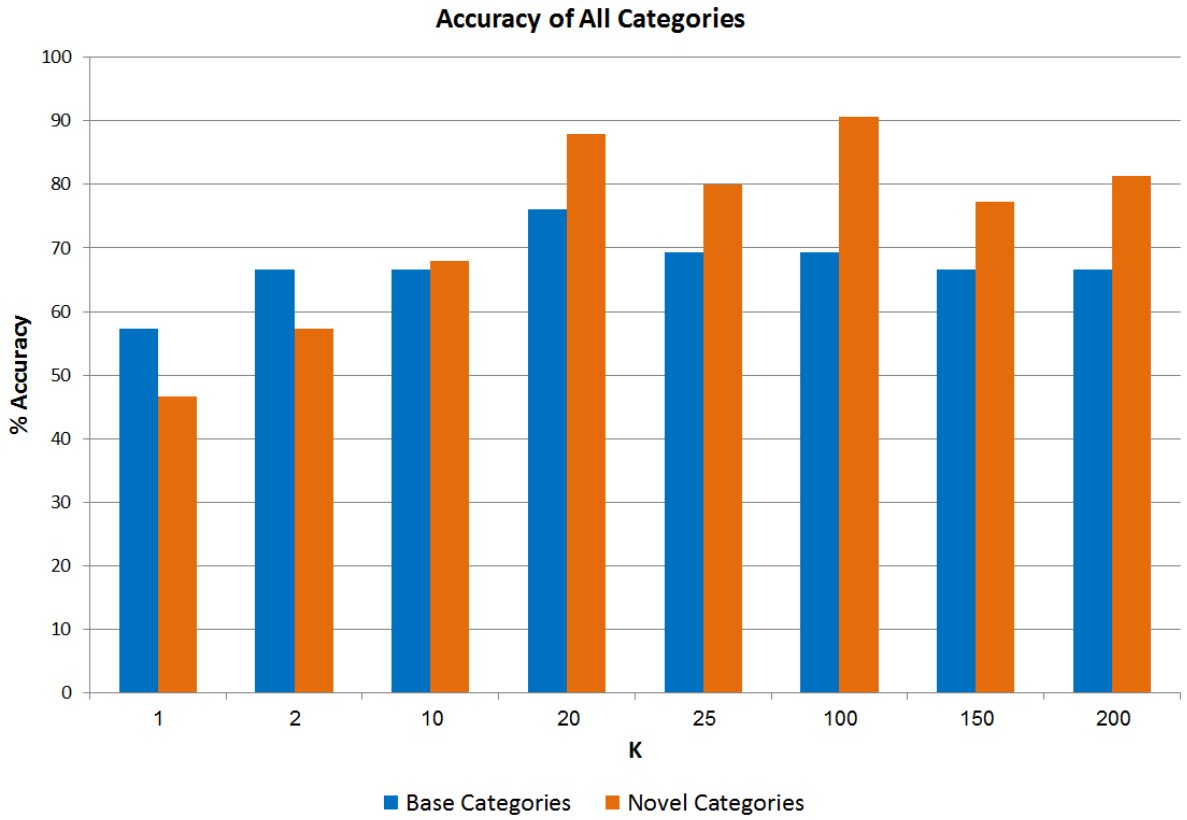


Figure 4.1: Accuracy results of *5-way k-shot* learning of Gidaris et al. [5], with a variety of k values, on base and novel categories. There is a trade-off between base and novel categories, which is dependent on the number of novel training samples, k. In cases of small values of k, the performance on base categories is higher than on novel ones, and both increase as k increases. With higher values of k, the accuracy on novel samples increases whereas there appears no improvement on base categories when k increases.

of very small training sets of the novel categories. We calculate the instability of the results, indicating the dependency on the number of shots, by standard deviation of all categories along different values of k, which is 9.25. Ideally, we would like the standard deviation to go to zero, indicating no dependency on the value of k.

4.4 Proposed Solution

In this section we propose a possible algorithm to avoid the base-novel trade-off presented in Section 4.3. Our goal is to reduce dependency of a few-shot learning model performance on the number of novel categories training samples, in order to extend its ability to perform in all real-life cases.

Algorithm 4.1 Feature Spaces Algorithm

1: Train the feature extractor module of Gidaris et al. [5].
2: Divide the base-category feature vectors evenly into N_s spaces.
3: **while** not done **do**
4: **for** each image **do**
5: Calculate the cosine-similarity score to the categories k in each space n and feature vector z:

$$S_{n_k} = \tau \frac{z^T}{||z^T||} \frac{W_{n_k}^*}{||W_{n_k}^*||}$$

where τ is a learnable scalar.
6: Find the maximal cosine-similarity score in each space $\left\{S_{n_{best}}\right\}_{n=1}^{N}$
7: Weighted calculation of the estimations to choose best category, and classifying the image accordingly.
8: **end for**
9: **end while**

4.4.1 Separated feature spaces

The main idea behind our approach is to decrease the imbalance between the number of base and novel training samples, since it results in the gap between the accuracy results of base and novel images. We propose dividing the extracted feature vectors of the trained base categories into separated feature spaces. Our approach also exploits the use of cosine-similarity function for generalizing better on the novel categories.

The algorithm is described in Alg. 4.1. The feature vectors of the base categories are divided evenly into N_s spaces. For classification, each image is matched to the best fitting category in each of the spaces separately. Then, using a weighted calculation, it combines the scores of matching this image in each of the N_s spaces in order to arrive to the best fit.

4.4.2 Experimental Results

We present a proof-of-concept for this proposed novel approach. There are many possibilities to divide the feature vectors into spaces, as well as many options of weighted calculation functions. In any case, the idea of diminishing the imbalance between base and novel categories, remains the same. In order to validate this approach we choose random separation for dividing the feature vectors into the spaces, and calculate a simple average of scores as the weighted calculation of spaces. The results of our experiments on *5-way k-shot* learning appear in Fig. 4.2.

Our method reduces the standard deviation of all categories along different values of k, from 9.25 to 6.5, meaning that the performance is less dependent on the value of k. Thus, the proposed model expresses higher stability then the one of Gidaris et al. [5] examined in Section 4.3. It can be seen from Fig. 4.2, that similar to the results of the original model, presented in Section 4.3, for small values of k the accuracy of novel categories classification improves as k increases. Moreover, our method improves the accuracy on base samples in those standard cases of few-shot learning with small values of k.

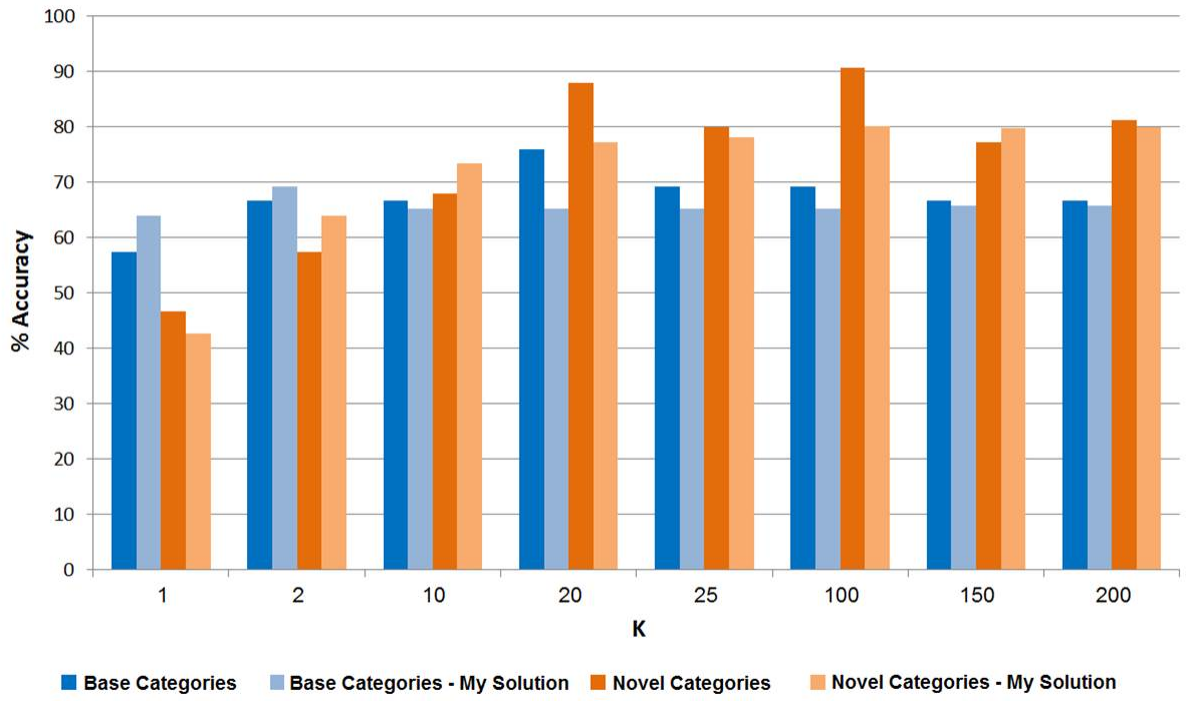


Figure 4.2: Accuracy results of *5-way k-shot* learning of the proposed method compared to Gidaris et al. [5], with a variety of k values, on base and novel categories. For small values of k our method improves the accuracy on base samples, and as k increases the accuracy on novel samples increases. For the non-standard cases of high k values, our model achieves better stability and less dependency on k.

For higher values of k, we observe that the performance of both, base and novel samples, remains almost stagnant. This implies a stability and diminished dependency on the number of novel training samples. Note that we only presented a proof-of-concept of the idea, training for no more than 20 iterations. Due to the performed stability, we believe that the proposed model has a very high potential. It may reach significantly higher accuracy on both categories when trained for more iterations.

4.5 Conclusions

We have investigated the trade-offs of few-shot learning algorithms, and proposed a solution to overcome them. First, we analyzed the accuracy results of the metric-learning combined with meta-learning mechanisms presented by Gidaris et al. [5] on mini-ImageNet dataset. Our settings comprised of 5 novel categories and several values of k, such that we can introduce the model with more shots than usually used in few-shot learning papers. We observed that for the standard few-shot learning cases, when k is small, the model performs better on novel categories than on base ones. As k increases, the accuracy on novel samples increases, whereas the accuracy on base samples does not rise. However with higher values of k the model acts differently and

there is a trade-off between base and novel performance. We note that the dependency on a small value of k is a great disadvantage of few-shot learning models in real-life applications, when there is no guarantee for a fixed small number of training samples.

Therefore, we identified a need to decrease dependency of base and novel performance on the value of k. We presented a novel approach to better absorb few novel categories in the feature spaces of base categories. The proposed method incorporates meta-learning with metric-learning approaches. It includes separation of the feature weight vectors of the base images into spaces and using a weighted calculations of scores in order to match a sample to the right category. We presented a proof-of-concept for this approach, which achieved a decreased standard deviation of the results along the different values of k. The results imply less dependency of the performance on the value of k. Moreover, the proposed method improved accuracy results for small values of k. We mention that more training procedures may contribute to accuracy improvements for all categories.

Another advantage of the proposed solution is that it is general. Thus, the same method can be implemented in many few-shot learning tasks, diminishing the imbalance between novel and base categories and decreasing the dependency on the value of k. In future work, we plan to further explore the separation into spaces, by examining the weighted calculation between matching scores in different spaces and possibly learning the best calculation. The number of spaces can also be optimized, as well as the method of separation into spaces. More work can be done on hyperparameters optimization, loss functions, non-linearity functions and training configurations.

Chapter 5

Conclusions

5.1 Research Summary

Audio-visual speech enhancement, although a well-studied problem, is still a major challenge for computers. In order to improve both quality and intelligibility, there is a need for previous knowledge about the noise or previous familiarity with the target speaker, a problem known as speaker dependency.

This thesis has introduced a novel approach for avoiding speaker dependency in audio-visual speech enhancement models. The proposed fast adaptation speech enhancement (FASE) model is inspired by methods of few-shot learning, originally developed for the task of image classification. FASE model outperforms previous art in the realistic cases where there is no guarantee for a previous familiarity with the target speaker.

Another significant benefit of FASE model is its small amount of parameters, which requires only small and accessible computational power. Thus, FASE model can be integrated in small and mobile systems of speech denoising, such as smart hearing aids.

The second part of the thesis involved with disadvantages and limitations of few-shot learning methods, examining their performance in the task of image classification. We discussed the trade-off between base and novel categories, and its dependency on the amount of novel training samples. We identified a need to decrease this dependency and suggested an algorithm to meet this need.

The results of our proof-of-concept exhibit a higher stability and a significant decreased variance of the performance between categories, which implies less dependency of the results on the amount of novel training samples. Furthermore, even though our solution was deployed in the task of image classification, it is general and can be implemented in many few-shot learning tasks, diminishing the imbalance between novel and base categories and the dependency on a fixed small number of shots.

5.2 Research Contributions

We can conclude the main contributions of our research as follows:

- **Avoiding speaker dependency in audio-visual speech enhancement tasks.** To the best of our knowledge, FASE model is the first meta-learning architecture to address the problem of speaker dependency in audio-visual speech enhancement. Our FASE algorithms demonstrates improved quality and intelligibility objective measurements. Moreover, FASE model exhibits improvements in compute time.
- **Extending few-shot learning methods to cope with more shots.** To the best of our knowledge, the proposed separation of the feature vectors into spaces is a unique approach to this problem. The proposed algorithm demonstrates better stability across different amounts of novel speaker training samples. Furthermore, our algorithm is general and therefore can be useful for many other few-shot learning tasks.

5.3 Future Research

The models presented in this thesis open several directions for future research:

- **Improving results of FASE model with very few shots.** Even though our FASE model managed to decrease speaker dependency and outperformed previous art in the cases of few-shot learning, still its PESQ and STOI measurements are not optimal. In order to improve the results, a larger dataset should be used and training configurations can be optimized (i.e learning rate and early stopping policies, number of layers in the network, etc.).

  Another idea is to learn a filter rather than the spectrogram itself. Learning a filter has been shown to improve quality and intelligibility of the output speech by Gabbay et al. in [27]. The filter is then multiplied with the original spectrogram in order to remove frequencies which do not relate to the visual features of the speech. The advantage of learning a filter is the inability to create sounds that do not exist in the original soundtrack. Thus, we can expect a fast convergence and the output would be cleaner.

- **Improving our few-shot learning model for more shots.** The algorithm presented in Chapter 4 improved the performance stability, but was only trained for 20 iterations as a proof-of-concept. Continuing the training procedure may contribute to accuracy improvements for all categories. More improvements can be achieved by changing the non-linearity functions, optimizing the learning rate and other training configurations, in order to better absorb the novel feature vectors in the separated feature spaces. Moreover, we plan to further explore the methods of separation into spaces, and to examine different methods for the weighted calculation between scores in different spaces. The number of spaces can also be optimized. For example, it can be dynamically chosen according to the number of base training samples.

- **Improving results of FASE model with more shots.** An interesting future work is finding an optimal model for audio-visual speech enhancement which avoids speaker dependency and achieves high performance both in the cases of few-shot learning and

in cases of more shots. Since the solution proposed in Chapter 4 is general and can be implemented in many few-shot learning tasks, we suggest deploying it in the task of audio-visual speech enhancement, for diminishing the trade-off between novel and base categories and the dependency on the value of k. In order to avoid speaker dependency and cope with both few and more shots, we propose to aggregate this algorithm together with FASE model and to train end-to-end.

Bibliography

[1] 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.

[2] Y. Kumar, M. Aggarwal, P. Nawal, S. Satoh, R. R. Shah, and R. Zimmermann, "Harnessing ai for speech reconstruction using multi-view silent video feed," in *Proc. 26th ACM International Conference on Multimedia*, 2018, pp. 1976–1983.

[3] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.

[4] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proc. Annual Meeting of the Cognitive Science Society*, vol. 33, no. 33, 2011.

[5] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4367–4375.

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[7] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2080–2094, 2012.

[8] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[9] G. Huang, J. Benesty, T. Long, and J. Chen, "A family of maximum SNR filters for noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2034–2047, 2014.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[11] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[12] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[13] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[14] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125–134, 2014.

[15] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, 2010.

[16] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *arXiv preprint arXiv:2008.09586*, 2020.

[17] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[18] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Proc. Interspeech*, 2013, pp. 436–440.

[19] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[20] Z. Chen, "Single channel auditory source separation with neural network," Ph.D. dissertation, Columbia University, 2017.

[21] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[22] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 2011 International Conference on Machine Learning (ICML)*, 2011.

[23] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Proc. Asian Conference on Computer Vision*. Springer, 2016, pp. 251–263.

[24] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[25] F. Khan and B. Milner, "Speaker separation using visually-derived binary masks," in *Proc. 2013 Auditory-Visual Speech Processing (AVSP)*, 2013, pp. 215–220.

[26] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[27] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Visually driven speaker separation and enhancement," in *ICASSP*. IEEE, 2018, pp. 3051–3055.

[28] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[29] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proc. Interspeech*. ISCA, 2018, pp. 1170–1174.

[30] T. Zhang and A. K. Bhowmik, "37-2: Invited paper: Enhancing speech in noisy and reverberant environments using deep learning techniques," in *Proc. SID Symposium Digest of Technical Papers*, vol. 49, no. 1. Wiley Online Library, 2018, pp. 467–470.

[31] J. Lee and H.-G. Kang, "A joint learning algorithm for complex-valued tf masks in deep learning-based single-channel speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1098–1108, 2019.

[32] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2015.

[33] J. Lee, J. Skoglund, T. Shabestary, and H.-G. Kang, "Phase-sensitive joint learning algorithms for deep learning-based speech enhancement," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1276–1280, 2018.

[34] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "Dnn-based source enhancement self-optimized by reinforcement learning using sound quality measurements," in *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 81–85.

[35] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5074–5078.

[36] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2016.

[37] Y. Kumar, R. Jain, M. Salik, R. ratn Shah, R. Zimmermann, and Y. Yin, "Mylipper: A personalized system for speech reconstruction using multi-view visual feeds," in *Proc. 2018 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2018, pp. 159–166.

[38] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. Advances in neural information processing systems*, 2018, pp. 4480–4490.

[39] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Video-driven speech reconstruction using generative adversarial networks," *arXiv preprint arXiv:1906.06301*, 2019.

[40] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[41] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," in *Proc. IEEE conference on computer vision and pattern recognition*, 2018, pp. 7278–7286.

[42] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, "A generative adversarial approach for zero-shot learning from noisy texts," in *Proc. IEEE conference on computer vision and pattern recognition*, 2018, pp. 1004–1013.

[43] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in Neural Information Processing Systems*, vol. 29, pp. 3630–3638, 2016.

[44] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4077–4087, 2017.

[45] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. International Conference on Machine Learning (ICML) Deep Learning Workshop*, vol. 2. Lille, 2015.

[46] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[47] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," *arXiv preprint arXiv:1711.04043*, 2017.

[48] A. S. Younger, S. Hochreiter, and P. R. Conwell, "Meta-learning with backpropagation," in *Proc. IJCNN'01. International Joint Conference on Neural Networks. (Cat. No. 01CH37222)*, vol. 3. IEEE, 2001.

[49] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[50] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

[51] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *arXiv preprint arXiv:1703.03400*, 2017.

[52] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," *arXiv preprint arXiv:1807.05960*, 2018.

[53] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," 2016.

[54] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. International Conference on Machine Learning*, 2016, pp. 1842–1850.

[55] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[56] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O Society*, vol. 12, no. 7, pp. 35–50, 1992.

[57] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1766–1776, 2007.

[58] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 625–638, 2009.

[59] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2299–2310, 2007.

[60] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007, vol. 615.

[61] J. Schmidhuber, "Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-hook," Ph.D. dissertation, Technische Universität München, 1987.

[62] S. Thrun and L. Pratt, *Learning to learn*. Springer Science & Business Media, 2012.

[63] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[64] A. Ephrat and S. Peleg, "Vid2speech: speech reconstruction from silent video," in *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5095–5099.

[65] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[66] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[67] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[68] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. 2010 IEEE International Conference on Acoustics, Speech and signal Processing*. IEEE, 2010, pp. 4214–4217.

[69] P. Mowlaee, R. Saeidi, M. G. Christensen, and R. Martin, "Subjective and objective quality assessment of single-channel speech separation algorithms," in *Proc. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 69–72.

[70] M. Anusuya and S. K. Katti, "Speech recognition by machine, a review," *arXiv preprint arXiv:1001.2267*, 2010.

[71] S. Bengio, Y. Bengio, J. Cloutier, and J. Gecsei, "On the optimization of a synaptic learning rule," in *Preprints Conf. Optimality in Artificial and Biological Neural Networks*, vol. 2. Univ. of Texas, 1992.

[72] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas, "Learning to learn by gradient descent by gradient descent," in *Proc. Advances in Neural Information Processing Systems*, 2016, pp. 3981–3989.

[73] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 3018–3027.

[74] H. Qi, M. Brown, and D. G. Lowe, "Low-shot learning with imprinted weights," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5822–5830.

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

הדוגמאות של הקטגוריות החדשות הזמינות לשלב האימון. לכן, אנחנו מציעים אלגוריתם לפתרון הבעיה ע"י חלוקת וקטורי המאפיינים למרחבים שונים. אנו מציגים הוכחת היתכנות של המודל, בה הושגה יציבות גבוהה יותר של התוצאות והופחתה התלות בכמות הדוגמאות לאימון. מעבר לכך, אנו מאמינים כי פוטנציאל רב טמון בגישה חדשה זו, כיוון שהיא כללית וניתנת ליישום גם במשימות אחרות של חוסר איזון בכמויות הדוגמאות לאימון.

התרומות העיקריות של עבודת תזה זו:

1. הפחתת תלות בדובר במשימת שיפור אות דיבור בוידאו - אנחנו מציעים רשת המתאימה עצמה במהירות לדוברים חדשים. הפתרון מבוסס רעיונות בהשראת גישת מטא-למידה שפותחו במקור למשימת למידה ממיעוט דוגמאות בסיווג תמונות. המודל מגיע לביצועים מרשימים במדדי איכות ומובנות של אות הדיבור של דובר חדש לו דוגמאות מעטות. מעבר לכך, המודל המוצע מפחית משמעותית את זמן הריצה ואת כוח החישוב הנדרש ועל כן בעל פוטנציאל עבור יישומים קטנים, כמו עזרי שמיעה.

2. הרחבת מודל למידה ממיעוט דוגמאות להתמודדות עם יותר דוגמאות - אנחנו מציעים שיטה חדשנית לשיפור מודלי למידה ממיעוט דוגמאות בסיווג תמונות, כך שביצועי המודלים לא יהיו תלויים בכמות הדוגמאות הקטנה. האלגוריתם המוצע מצטרף למרחב המאפיינים היוצאים מהרשת. על כן הוא כללי ויוכל להתאים למגוון של בעיות למידה ממיעוט דוגמאות.

להערכתנו שילוב שני הרעיונות שפותחו בעבודה זו יתרום ליצירת מודל שיפור אות דיבור אופטימלי המתאים למקרים מציאותיים של מיעוט ידע קודם, שאינו תלוי בדובר וכן אינו תלוי בכמות הדוגמאות ללמידה.

בעבודת תזה זו אנו עוסקים בבעיה של תלות בדובר. אנחנו בוחנים את המקרה המציאותי בו קיימת רק כמות קטנה של דוגמאות של הדובר לאימון המודל. לשם כך, אנו מציעים רשת שיפור אות דיבור המתאימה את עצמה במהירות לדוברים חדשים. את ההשראה למודל קיבלנו משיטות של למידה ממיעוט דוגמאות שפותחו במקור עבור משימת סיווג תמונות. זאת, לאור הדימיון בין בעיית התלות בדובר לבעיית המחסור בתמונות מקטגוריה מסויימת לאימון. בשני המקרים אין כמות מספקת של דוגמאות ללמידה של חלק מהאפשרויות.

למידה עמוקה ממיעוט דוגמאות הוא נושא משמעותי במשימת סיווג תמונות. לעומת בני אדם המצליחים ללמוד קטגוריה חדשה ממעט דוגמאות כבר בגיל צעיר, מרבית אלגוריתמי הלמידה המפוקחת נדרשים להתאמן על כמות גדולה של תמונות כדי להגיע לביצועים דומים. לשם כך נדרש לאסוף את המידע למאגר נתונים מגוון ואיכותי וכן לתייג את התמונות. במקרים רבים עולה קושי לעשות זאת, הן מסיבות כספיות, הן בעקבות זמני פיתוח והן מכיוון שבבעיות סיווג לעיתים יש קטגוריות שבהן לא ניתן לצלם מספיק דוגמאות. במקרים רבים, הקטגוריות הנדירות הן דווקא החשובות ביותר לזיהוי על ידי המערכת. בנוסף לכך, לאחר שלב האימון של מודל למידה עמוקה, הדורש לרוב זמן רב וכוח חישוב גדול, לא ניתן להוסיף או ליצור התאמה של המודל לקטגוריות חדשות בקלות. שיטות למידה ממיעוט דוגמאות עוזרות להתמודד עם בעיה זו.

האתגר בשיטות למידה ממיעוט דוגמאות הוא חוסר האיזון בין כמות דוגמאות האימון בקטגוריות השונות, שעלול לגרום להטיה בתוצאות הסיווג של המודל. ניתן לחלק את שיטות ההתמודדות עם בעיה זו לשלוש גישות עיקריות: שיטות אוגמנטציה, המייצרות דוגמאות חדשות מהקיימות; שיטות למידת מטריקה, המשתמשות במרחב המאפיינים על מנת ליצור קרבה בין דוגמאות של כל קטגוריה והגדלת ההפרדה בין קטגוריות שונות; ושיטות מטא-למידה, בהן הרשת לומדת את הפרמטרים שיעזרו ללמוד את יכולת ההתמודדות עם קטגוריות חדשות. זאת באמצעות יצירת איטרציות אימון, שבכל אחת מהן התמודדות עם הבעיה של למידת קטגוריה אחרת בה יש מיעוט דוגמאות.

מודל שיפור אותות הדיבור שאנו מציעים עבור התמודדות עם בעיית התלות בדובר, נוצר מתוך השראה משיטות מטא-למידה. הוא מכיל רשת עמוקה של מקודד ומפענח, המאומנים על ידי רשת חיצונית עבור אדפטציה מהירה לדוברים חדשים. למעשה, המודל מתבסס על לימוד איטרטיבי של רשת המקודד והמפענח כיצד להתמודד עם הבעיה של אדפטציה מהירה לדוברים חדשים. בכל איטרציה של הרשת החיצונית, המודל צריך להתמודד עם בעית המחסור במידע של דובר אחר. בכך למעשה הרשת החיצונית מאמנת את הרשת הפנימית של המקודד והמפענח על משימות שונות של מיעוט דוגמאות.

ככל שידוע לנו, מודל המשלב שיטות מטא-למידה למניעת בעיית התלות בדובר, הינו חדש ופורץ דרך בתחום שיפור אות דיבור. אנו מציגים תוצאות של המודל במגוון מקרים של כמות דוגמאות של הדובר החדש הזמינה לאימון. בכל המקרים של מיעוט משמעותי של דוגמאות, המודל המוצע עולה בביצועיו על מודלים קודמים ואכן מפחית את הבעיה של תלות בדובר. מעבר לכך, המודל המוצע מכיל כמות קטנה של פרמטרים ודורש כוח חישוב קטן יותר. לכן, הוא בעל יתרון ביישומים ניידים ולבישים.

בהמשך המחקר אנחנו מתמקדים בחסרונות של רשתות למידה ממיעוט דוגמאות. תחילה אנו בוחנים את המודל שהצענו עבור שיפור אות הדיבור. ראינו כי ביצועיו מרשימים במקרים בהם כמות הדוגמאות של הדובר החדש נמוכה במיוחד. לעומת זאת, כאשר כמות הדוגמאות גדלה, המודלים הסטנדרטיים מבוססי התלות בדובר משיגים ביצועים טובים יותר. לפיכך, אנחנו מחליטים לבחון לעומק את המגבלות של רשת למידה עמוקה ממיעוט דוגמאות במשימה המקורית של סיווג תמונות.

מתוצאות הבדיקה ניתן להבחין בחוסר יציבות של התוצאות, הנובע מתלות של מודל הלמידה העמוקה בכמות

תקציר

הקלטה של אות דיבור בסביבה טבעית לרוב כוללת רעש רקע. הרעש יכול להיווצר מהסביבה, מדוברים אחרים ברקע וכן ממכשיר ההקלטה עצמו. במשך מספר עשורים, פותחו מודלים ממוחשבים לשיפור אות דיבור. מודלים אלו מקבלים כקלט את אות הדיבור המורעש ומוציאים פלט בו אות הדיבור של הדובר משופר ואילו רעשי הרקע מונחתים. לאחרונה, יחד עם אות הדיבור המועבר למודל, החלו להעביר גם את רצף התמונות המתאימות מסרטון הוידאו. הוספת לבין המידע החזותי, כדוגמת תנועת השפתיים ושפת הגוף, שיפרה את יכולות האלגוריתמים שהיו מבוססי אות דיבור בלבד. זאת לאור הקשר בין המידע החזותי לאות הדיבור של הדובר, ללא תלות ברעשי הרקע.

שיפור אות דיבור על סמך מידע קולי, או מידע קולי וחזותי יחד, קריטית ליישומים רבים בזמן-אמת. בין היתר ניתן למנות:

1. עזרי שמיעה חכמים - מודל שיפור אות דיבור מבצע עיבוד מקדים לאותות דיבור הנקלטים במכשיר שמיעה. לאחר שיפור הדיבור והפחתת הרעש, מוגבר האות הנקי עבור מוגבל השמיעה.

2. העברת אותות תקשורת - בתקשורת קולית קיימת בעיה של רעשי רקע המגיעים לעמדת הקצה. מודל שיפור אות דיבור יכול לשמש גם ביישומים אלו על מנת להפחית את הרעש ולשפר את איכות הקול לפני הגעתו מהמשדר למקלט.

3. יצירת כתוביות אוטומטיות לסרטים - במערכות כתוביות אוטומטיות עולה קושי כאשר כמה דוברים מדברים בו זמנית. הפרדת הדוברים בעזרת אלגוריתם שיפור אות דיבור יכולה להקל על יצירת כתוביות מדוייקות יותר עבור כל דובר.

4. ניתן למנות מגוון יישומים נוספים בתחומי המדיה, עוזרים וירטואליים, פרסום מותאם אישית, יישומי מודיעין ויישומים רפואיים.

לעומת בני אדם, עבור מודלים ממוחשבים, שיפור אות דיבור ללא כל היכרות מקדימה עם הדובר, מהווה אתגר משמעותי. בעיה זו נקראת - תלות בדובר. מודלים קלאסיים לשיפור אות דיבור נאלצו להניח הנחות אודות מאפייני הדובר הספציפי או הנחות אודות הרעש. רמת ביצועיהם, הנמדדת על ידי מדדי איכות ומדדי מובנות אובייקטיביים של הדיבור, היתה תלויה בהנחות אלו. לאחרונה, גברה ההתעניינות ביצירת מודלים מבוססי למידה עמוקה עבור שיפור אות דיבור בוידאו. אלגוריתמי הלמידה אכן הגיעו לביצועים מרשימים וגברו על חסרונות האלגוריתמיים הקלאסיים, אך עדיין ביצועיהם מוגבלים ותלויים בלמידת דובר מסויים. בפועל, מודל למידה עמוקה נדרש לקבל בשלב האימון כמות גדולה של דוגמאות דיבור של הדובר המסוים אותו עליו להגביר במהלך שלב המבחן. בעיית התלות בדובר מונעת מאלגוריתמי שיפור אות דיבור לפעול ביישומים בזמן-אמת, בהם אין כל הבטחה כי יהיו מספיק דוגמאות של הדובר לאימון המודל.

המחקר בוצע בהנחייתו של פרופסור ישראל כהן, בפקולטה להנדסת חשמל ומחשבים.

חלק מן התוצאות בחיבור זה נשלחו לפרסום כמאמר מאת המחברת והמנחה ב-

IEEE/ACM Transactions on Audio, Speech, and Language Processing.

תודות

תודה למנחה שלי, פרופ' ישראל כהן.

תודה למשפחה ולחברים שתמכו לאורך הדרך.

.

רשת נוירונים למיעוט דוגמאות עבור שיפור אות דיבור בווידאו

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים בהנדסת חשמל ומחשבים

מאיה רפפורט

הוגש לסנט הטכניון — מכון טכנולוגי לישראל
אייר התשפ"א חיפה מאי 2021

רשת נוירונים למיעוט דוגמאות
עבור שיפור אות דיבור בווידאו

מאיה רפפורט