UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Audio-Visual Speech Separation

Diploma Thesis

Apostolos Karasmanoglou

Supervisor: Assoc. Prof. Gerasimos Potamianos

Volos, September 2021

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Audio-Visual Speech Separation

Diploma Thesis

Apostolos Karasmanoglou

Supervisor: Assoc. Prof. Gerasimos Potamianos

Volos, September 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Οπτικό-ακουστικός Διαχωρισμός Φωνής

Διπλωματική Εργασία

Καρασμάνογλου Απόστολος

Επιβλέπων: Αναπλ. Καθ. Γεράσιμος Ποταμιάνος

Βόλος, Σεπτέμβριος 2021

Approved by the Examination Committee:

Supervisor **Gerasimos Potamianos**
Associate Professor,
Department of Electrical and Computer Engineering,
University of Thessaly

Member **Antonios Argyriou**
Associate Professor,
Department of Electrical and Computer Engineering,
University of Thessaly

Member **Dimitrios Katsaros**
Associate Professor,
Department of Electrical and Computer Engineering,
University of Thessaly

Acknowledgements

First and foremost, I would like to thank my supervisor, Associate Professor Gerasimos Potamianos for his guidance and patience despite my occasional lateness and infrequent communication, and for playing an important role in introducing me to Signal Processing and Pattern Recognition, greatly influencing my academic interests.

Secondly, I would like to thank all professors and teaching staff of the Electrical and Computer Engineering Department at the University of Thessaly, who throughout my years as an undergraduate have guided me towards demystifying computer science in their respective fields, helping me conquer ideas that would inspire awe in me years ago.

I feel obliged to express my gratitude for anyone who has laboured emotionally to the benefit of my sanity during this trying year, hardened by the challenge of completing this Thesis: my friends, family, and loved ones. This Thesis would have never materialized without their support.

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

«Being fully aware of the implications of copyright laws, I expressly state that this Diploma Thesis, as well as the electronic files and source codes developed or modified in the course of this Thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this Thesis or part of it does not belong to me because it is a product of plagiarism».

The declarant

Apostolos Karasmanoglou

Abstract

Speech signal separation involves separating the component speech signals from an audio mixture in which they are combined. In the field of independent component analysis, this problem is known as the ”cocktail party problem” that describes the inherent human ability to isolate specific auditory stimuli in noisy environments. For example, we encounter the cocktail party phenomenon when focusing our attention to listen to a friend talking in a crowded cafe, a busy street, or a bar with loud music playing. In audio-visual speech separation (AVSS), information from both audio and visual data is used in order to take advantage of the multi-modal nature of human speech in the separation procedure. Recent research in AVSS algorithms employs Deep Neural Network architectures to map the two modalities to a latent vector space, and subsequently fusing them to extract the underlying speech signals. The goal of this Thesis is to design our own architecture for the AVSS task. In addition, we evaluate the separation performance of the proposed architecture on data from the Lombard GRID and TCD-TIMIT databases under various experimental conditions designed to model realistic scenarios of speech signal separation.

Περίληψη

Ο διαχωρισμός σήματος ομιλίας περιλαμβάνει τον διαχωρισμό των συστατικών σημάτων ομιλίας από μια ηχητική ανάμειξή τους στην οποία συνδυάζονται. Στον τομέα της ανάλυσης ανεξάρτητων συνιστωσών αυτό το πρόβλημα είναι γνωστό ως ”πρόβλημα κοκτέιλ πάρτυ”, το οποίο αναφέρεται στην εγγενή ικανότητα των ανθρώπων να απομονώνουν συγκεκριμένα ακουστικά ερεθίσματα σε θορυβώδη περιβάλλοντα. Για παράδειγμα, συναντάμε το φαινόμενο του πάρτυ κοκτέιλ όταν εστιάζουμε την προσοχή μας για να ακούσουμε ένα φίλο να μιλά σε ένα γεμάτο καφέ, σε έναν δρόμο με έντονη κίνηση ή σε ένα μπαρ με δυνατή μουσική. Στον οπτικοακουστικό διαχωρισμό ομιλίας, πληροφορίες τόσο από ακουστικά όσο και από οπτικά δεδομένα χρησιμοποιούνται προκειμένου να επωφεληθεί η διαδικασία του διαχωρισμού από την πολύτροπη φύση της ανθρώπινης ομιλίας. Πρόσφατες μελέτες σε αλγορίθμους οπτικοακουστικού διαχωρισμού ομιλίας χρησιμοποιούν αρχιτεκτονικές βαθιών νευρωνικών δικτύων για την απεικόνιση των δεδομένων σε έναν λανθάνοντα διανυσματικό χώρο και επακόλουθα για τη σύνθεσή τους προς την εξαγωγή των υποκείμενων σημάτων ομιλίας. Ο στόχος αυτής της διπλωματικής εργασίας είναι να σχεδιάσουμε μια δικιά μας αρχιτεκτονική για αυτήν τη λειτουργία. Επιπλέον, εξετάζουμε την αποτελεσματικότητα διαχωρισμού της προτεινόμενης αρχιτεκτονικής σε δεδομένα από τις βάσεις Lombard GRID και TCD-TIMIT υπό διάφορες πειραματικές συνθήκες, σχεδιασμένες ώστε να μοντελοποιούν ρεαλιστικά σενάρια διαχωρισμού σημάτων φωνής.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Seeing Speech?

In the early days of air traffic control, operators would receive messages from pilots over a central loudspeaker, instead of individual headsets. Often these messages would be intermixed, with many pilots talking over one another, so operators would have to focus their attention on specific pilot voices in order to decipher a message. Based on this observation, in 1953 cognitive scientist C. E. Cherry would coin the ”cocktail party effect”: our inherent ability to ”tune in” to specific speech signals and ”tune out” others in a noisy environment. Cherry would then remark on the complexity of the cognitive functions that make this possible. To illustrate, he listed the aspects of speech that an ”automatic voice recognition machine” could exploit in order to perform the same task: lipreading, accent differences, different directions of the voice source, differences in gender, etc [1]. Later he would note [2]:

”One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may take it for granted; we may call it “the cocktail party problem.” No machine has been constructed to do just this, to filter out one conversation from a number jumbled together.” - Collin E. Cherry 1957

In recent years, ”cognitive machines” in the form of neural networks have been dominating the field of artificial intelligence including speech applications. The goal of designing the machine proposed by Cherry seems even more feasible today, given the success of these models. Furthermore, the cocktail party problem has been thoroughly formalized mathematically,

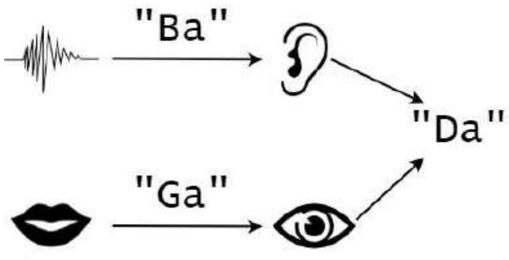


Figure 1.1: Illustration of the McGurk effect

and many approaches have been tested to tackle this problem in real-world applications.

Furthermore, the trend of user speech interfaces in personal assistants in recent years such as Apple’s Siri or Amazon’s Alexa, as well as several augmented reality applications, have created the incentive to further develop the field of automatic speech recognition. In typical real-world usage of such assistants, the need for de-noising speech signals in acoustically complex environments arises naturally. Methods for audio-visual speech separation (AVSS) may prove useful here: removing interfering speech from user input is by far one of the most challenging cases of noise suppression, and it may require the use of a camera in order for it to be successfully performed.

However, incorporating computer vision in automatic speech recognition is not merely a high-concept trick for improving performance: speech interpretation might be fundamentally incomplete without the visual mode. Going back to Cherry’s original statement, a clear emphasis on the audio-visual aspect of source separation is placed. Humans leverage the multi-modal nature of speech in order to perceive speech in noisy environments, employing lipreading when audio-only perception is insufficient. This fact is illustrated by the ”McGurk effect” [3], an illusion in which the perceived audio emanating from a speaker is transformed by the visual stimuli that it is paired with. For example: if a ”ba” sound produced by a speaker is accompanied by visuals of the facial movements associated with the production of a ”ga” sound, the perceived speech will be ”da”.

From the above it becomes clear that, as speech recognition systems get more and more sophisticated, the inclusion of a visual stream in speech recognition algorithms is inevitable, and thus research on methods for manipulating audio-visual speech is a necessity.

1.2 Thesis Goals and Contributions

Our goal is to implement an AVSS system using a multi-modal fusion mechanism for combining temporally aligned audio and video data in order to infer relationships between the two. The inferred information will then be used in decoding each speaker signal and reconstructing the separated sources. More specifically, we list our main goals for our approach:

- Speaker Independence: Some speech separation architectures are designed in a speaker-dependent fashion in order to exploit a certain speaker's distinct vocal characteristics for separation. Our proposed model instead is trained on datasets that have been split in such a way so that the set of test speakers is disjoint from the set of training speakers. This way, our evaluation of the model on the test set ensures generalization to separating speakers that the model has no prior knowledge of.

- Scalability: A common problem amongst speech separation architectures is that they are often designed for multi-regression purposes. In these, a corresponding speech signal is regressed for each speaker in the mixture. For this to work, the total number of parameters used by the visual network must be increased as more and more speakers are added. We design our model in such a way, as to not need to increase these parameters, instead applying a one-against-the-rest distinction of the visual modalities, in which one targeted speaker's visual modality is simply contrasted with all others.

- Applicability under different noise conditions: We train and evaluate our model under different interference amplitude conditions, from cases where the noise is very low to very adverse acoustic environments where the noise may surpass that of the target speaker. Good separation performance on all levels of noise is desired, with the higher noise conditions being more challenging.

The contributions of this Thesis to the field of AVSS are:

- A thorough exploration of the formal background of speech signal separation, reviewing methods used in Blind Source Separation (BSS) and Independent Component Analysis (ICA), as well as how they are applied to AVSS.

- The development of a U-Net style speech separation network architecture, consisting of strictly convolutional layers, with a unique audio-visual fusion gating mechanism for visually guided speech separation. The proposed model is designed in accordance to the goals outlined above.

- An evaluation of our proposed architecture for different speech separation tasks and speaker gender pairings on two appropriate datasets.

1.3 Related Work

Recently, Gao [4], Ephrat, [5] and Afouras [6] have developed some of the most impressive speech source separation architectures, capable of separating unseen speaker mixtures from real-life data with remarkable efficiency. Demos of these systems can be found on YouTube showcasing impressive separation results. The architecture of [4] is very similar to the one we propose, consisting of a U-Net audio analysis network that is fused in the middle part with the outputs from a lipreading 3D convolutional network, as well as a facial attribute analysis network that leverages correlation of certain speaker facial features with aspects in the corresponding speech (for example, male facial features correlate with lower pitch voices). The model proposed by Ephrat et al. [5] is an extension of earlier work [7] that successfully furthers the capability of this architecture to separating unknown speakers, namely speakers that have not been seen during model training. Their model extracts features from audio and visual streams using a sequence of dilated convolutional layers, concatenating the resulting features and passing them through a bi-directional long short-term memory model (LSTM) to capture sequence correlations. This architecture has proven successful in separating speech signals in environments with background noise present. A similar approach is presented in [6], extracting features from the two streams via convolutional layers, concatenating them and passing them through a fully-connected layer. What is interesting about this architecture is that first, only the magnitude part of each signal is separated, and subsequently the resulting signals are phase corrected using a separate module designed for estimating the appropriate phase correction. Notably, all aforementioned approaches estimate a spectral mask that is used for separating the speech signals. We adopt this approach for our proposed model as well, and we describe it in detail further in this Thesis.

A different approach to the AVSS problem is presented in [8], where audio and visual embeddings are extracted from mixture audio spectrogram segments and the corresponding speaker video frames using a combination of convolutional layers and bi-directional LSTMs.

The embeddings are subsequently fused, with the resulting audio-visual embeddings used for deep clustering of the spectral components present in the audio mixture. The resulting clusters correspond to the separated speech signals. Also, another recent publication [9] proposes the use of variational autoencoders to separate speech mixtures.

Our model takes inspiration from all previously cited approaches, as well as the multimodal transform module presented in [10], used for speech enhancement, from which the inspiration of the squeeze-excite fusion mechanism [11] that we employ is taken and modified to suit our needs.

Finally, note that the U-Net architecture we will be using for speech separation in the time domain has been employed in other similar tasks, such as audio-visual music source separation [12], singing voice separation [13], and speech enhancement [14].

1.4 Thesis Outline

The remainder of this Thesis consists of the following four Chapters:

- Ch 2. Background: First, we review the fundamentals of a mathematical model of the cocktail party problem. Beginning from rudimentary models for separating simple linear mixtures of signals, we build up to the spectral masking technique for separation in the time-frequency domain, which is the central objective of our model.

- Ch 3. Implementation: Following, we present our design rationale for the speech separation architecture we designed. We introduce each of its individual parts step by step, showing the utility of each one in the separation pipeline.

- Ch 4. Results: Next, we present the performance evaluation of our model, testing it for different types of speech separation tasks. We measure separation performance using several different metrics, contrasting the efficiency of our approach to other related architectures.

- Ch 5. Conclusion: Finally, we conclude this Thesis by summarizing our evaluation findings, noting our proposed model's advantages as well as its limitations, and discussing our ideas for how to potentially improve and test our design further.

Chapter 2

Background

In this chapter we explore the mathematical descriptions of the cocktail party problem used by methods for BSS and ICA (Sections 2.1,2.2, and 2.3). We first introduce core concepts of these disciplines, embellishing them eventually leading to the concept of spectral masks (Section 2.4). Our end goal is to clarify the motivation behind choosing a spectral mask as the target for the model to approximate. Finally, we discuss how spectral masks can be learned within a deep learning framework, also incorporating the visual modality (Section 2.5).

2.1 A Simple Source Separation Framework

We begin by introducing some core principles commonly used in ICA and BSS to describe the cocktail party problem mathematically. The two fields are closely related, with their key difference being that in BSS the goal is to extract statistically independent signals combined via some unknown procedure that is assumed to be a linear transformation, whereas ICA provides a probabilistic tool for dealing with mixtures of independent components (random variables), thought to be combined linearly [15].

Let us consider a set of discrete, real-valued signal components $\{s_i\}_{i=1}^{N}$, each adhering to one of N signal sources. These signals are combined by some unknown process, yielding a set of observed signals $\{x_j\}_{j=1}^{M}$ at M different destinations, where measuring devices are installed. The standard framework for modeling signal mixtures is a simple linear combination of the signal components at each time frame n:

$$x_j[n] = \sum_{i=1}^{N} a_{ji} s_i[n] \ , \quad \text{for} \ \ j = 1, \ldots, M \ . \tag{2.1}$$

In order to model random noise in the j-th measurement device, we add the component u_j to the mixture. This additive noise component is uncorrelated with the source signals and ”white”, i.e. any two noise signals u_j, u_i from different measuring devices $i \neq j$ are uncorrelated with each other, and also any two values of a noise signal at different time frames are uncorrelated with each other. Thus:

$$x_j[n] = \sum_{i=1}^{N} a_{ji} s_i[n] + u_j[n] \ , \quad \text{for} \ \ j = 1, \ldots, M \ , \tag{2.2}$$

or using a more compressed notation:

$$\vec{x}[n] = A\vec{s}[n] + \vec{u}[n] \ , \tag{2.3}$$

where $A \in \mathbb{R}^{M \times N}$ is a matrix of mixture parameters and $\vec{x}[n]$, $\vec{s}[n]$ are vectors consisting of the component values in the n-th time frame:

$$\begin{aligned} \vec{x}[n] &= \begin{bmatrix} x_1[n] & x_2[n] & \ldots & x_M[n] \end{bmatrix}^T , \\ \vec{s}[n] &= \begin{bmatrix} s_1[n] & s_2[n] & \ldots & s_N[n] \end{bmatrix}^T , \\ \vec{u}[n] &= \begin{bmatrix} u_1[n] & u_2[n] & \ldots & u_M[n] \end{bmatrix}^T . \end{aligned}$$

The goal of ICA estimation is to derive the original source signals reconstructed using the mixed data observations $\vec{s}[n]$. This is done by de-noising observations and then estimating the unmixing matrix, i.e. the inverse of matrix A denoted as $W = A^{-1}$. The reconstructed estimation is derived by the following generative model, after omitting noise:

$$\hat{s}[n] = \hat{W}\vec{x}[n] = A^{-1}\vec{x}[n] \ , \tag{2.4}$$

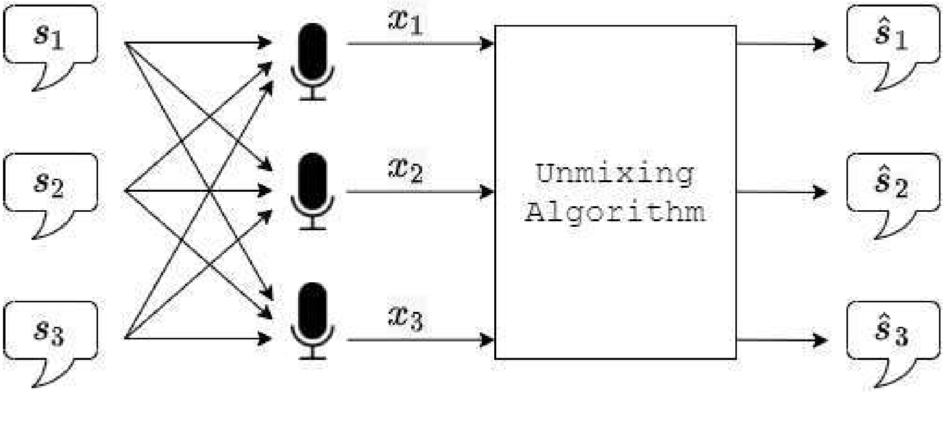


Figure 2.1: ICA/BSS view of the cocktail party problem for 3 speakers

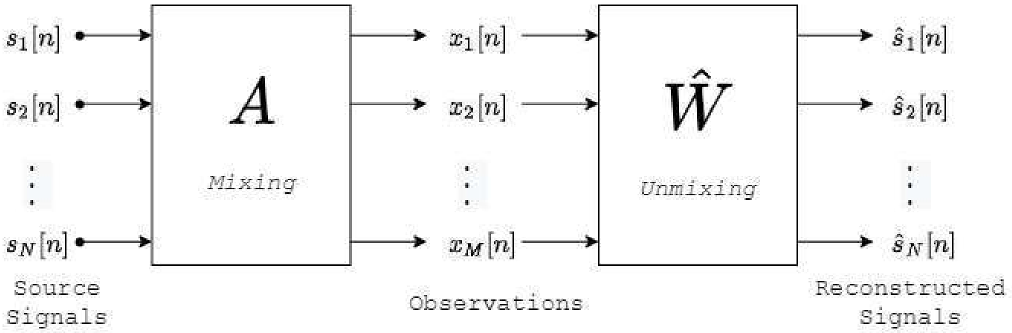


Figure 2.2: Mixing and unmixing matrices in BSS

as in [16, 17], where $\hat{W}$ is an approximation of the ground-truth unmixing matrix.

A key assumption of ICA is the independence of the variables that are being combined, something that cannot be practically guaranteed, however we can inspect the uncorrelatedness of the variables, a slightly less stringent constraint than independence. Generally, two real random variables x, y are uncorrelated if:

$$\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] = 0 \;, \tag{2.5}$$

whereas independence supposes that for every real function $h : \mathbb{R} \to \mathbb{R}$ the following holds:

$$\mathbb{E}[h(x)h(y)] - \mathbb{E}[h(x)]\mathbb{E}[h(y)] = 0 \;. \tag{2.6}$$

This is the independence constraint assumed in classical ICA/BSS [18, 19].

Having outlined the process of ICA estimation, it is useful to comment on some of its main ambiguities encountered when unmixing independent components in this fashion [18]

- Permutation: Because ordering of the observations $x_1, x_2, \ldots, x_M$ is arbitrary and $A, \vec{s}[n]$ are unknown parameters of the generative model, any ordering of the reconstructed sources is in essence valid. Post-separation sorting needs to be applied, if a specific order in the extracted signal components is expected.

- Scaling or variation: The power or variation of the signal may vary from source to source, and, due to our lack of knowledge of the structure of A, an arbitrary scalar multiplier applied to a row of A will distort the power level of the reconstructed signals. A common assumption made to remedy this is that the power level of each component is normalized: $\mathbb{E}\left[s_i^2[n]\right] = 1$. After unmixing, the reconstructed components can be amplified accordingly.

- Under-determinedness: It is not rare that the number of sources in a mixture is larger than that of the available measuring devices. As stated in [20], any number of M, N destination devices and source signals may be encountered in BSS applications, yielding under-determined, over-determined, or properly determined linear systems for the generative model. Among these cases, mono-aural source separation refers to the task of separating signals sampled from a single device, i.e. $M = 1$.

Having discussed these ambiguities, we re-write the formerly defined task of ICA estimation, so that it may be described as a three-part transformation process of the observation sequences, each corresponding to a matrix:

$$\hat{s}[n] = \Lambda P W \vec{x}[n] \ , \tag{2.7}$$

as in [21]. Each of the matrices Λ, P, W serves a specific purpose in addressing a kind of aforementioned ambiguity. Specifically:

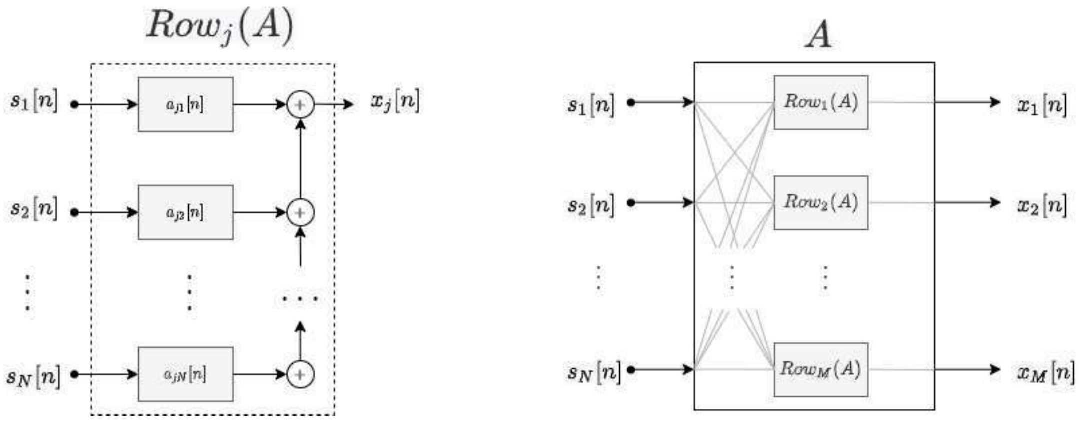
- $\Lambda \in \mathbb{R}^{N\times N}$ is a diagonal matrix with its elements set to amplify or diminish the reconstructed signal energy to that appropriate for the corresponding source.

- $P \in \{0,1\}^{N\times N}$ is a permutation matrix reversing any incorrect permutation in the assignment of extracted signals to sources.

- $W \in \mathbb{R}^{N\times M}$ is the unmixing matrix introduced earlier, restated as the generalized inverse of A, i.e. $W = A^{\dagger}$, meaning that a pseudo-inverse is used when the system is under-determined to derive an optimal estimation of the ground-truth signals.

It is easy to see how the ICA/BSS framework incorporates the cocktail party problem. In a complex auditory environment, N audio sources (multiple speakers, background music, traffic, ambient sounds, etc.) $\{s_i[n]\}_{i=1}^{N}$ may be combined linearly into a mixture, and measured by M different devices More specifically, in our application of interest we are concerned with a mono-aural speech mixture, where multiple speech signals are superimposed linearly:

$$x_{mix}[n] = \sum_{i=1}^{N} a_i s_i[n] \ . \tag{2.8}$$



(i) Signal mixing model for a single device (ii) Signal mixing model for an array of devices

Figure 2.3: Convolutional mixing BSS

2.2 Convolutional Source Separation

The basic BSS model described by equations (2.1) and (2.2) can be thought of as an ”instantaneous mixing model”, where any one of the observed variables $x_j[n]$ at time step n is only affected by the values of the source signals at the given instance $s_i[n]$. The motivation of extending the definition of BSS to convolutional mixtures arises from the need to model more complex natural mixing environments, where the signal component samples within a given time window of length K, $\{s_i[n+k]\}_{k=0}^{K-1}$ are weighted and delayed in time, thus contributing to the observation at the destination in a time-distributed fashion. This sort of function can be described by a convolutional mixture model [22], extending (2.1) as:

$$x_j[n] = \sum_{i=1}^{N}(a_{ji} * s_i)[n] = \sum_{i=1}^{N}\sum_{k=0}^{K-1} a_{ji}[k] \cdot s_i[n-k] \ , \tag{2.9}$$

where symbol $*$ denotes the convolution operation, or equivalently in matrix notation:

$$\vec{x}[n] = (A * \vec{s}\ )[n] = \sum_{k=0}^{K-1} A[k]\vec{s}[n-k] \ , \tag{2.10}$$

as in [23]. We can once again add a white noise component to this model to account for noise in measurements:

$$\vec{x}[n] = (A * \vec{s}\ )[n] + \vec{u}[n] \ , \tag{2.11}$$

but we omit this from our formulation for now.

As one can observe, the mixing matrix is no longer a flat 2D matrix, but each component a_{ij} is an FIR filter that contains a set of K convolution coefficients distributed in time $a_{ij} \in \mathbb{R}^K$, and A is said to be an FIR matrix [24].

Within this framework, the problem of ICA estimation is restated as a deconvolution problem, i.e. the estimation of an unmixing filter matrix $W \in \mathbb{R}^{K \times M \times N}$ capable of reconstructing the source signals when convolved with the observation signal vector:

$$\hat{s}[n] = (\hat{W} * \vec{x}\,)[n] = \sum_{k=0}^{K-1} \hat{W}[k]\vec{x}[n-k] \; . \tag{2.12}$$

Convolutional mixing of source signals can be better understood in the frequency domain, as the convolution theorem makes computations much simpler. By applying a Discrete Fourier Transformation (DFT):

$$\vec{s}[n] \xrightarrow{DFT} \vec{S}[f] \; ,$$

$$\vec{x}[n] \xrightarrow{DFT} \vec{X}[f] \; .$$

Mixture and unmixture of components can thus be expressed in the frequency domain:

$$\vec{X}[f] = A[f]\vec{S}[f] \; , \tag{2.13}$$

$$\hat{S}[f] = \hat{W}[f]\vec{X}[f] \; . \tag{2.14}$$

Adding a noise component to this formulation i.e. $\vec{u}(t) \xrightarrow{DFT} \vec{U}[f]$, yields the equations:

$$\vec{X}[f] = A[f]\vec{S}[f] + \vec{U}[f] \; , \tag{2.15}$$

$$\hat{S}[f] = \hat{W}[f]\vec{X}[f] + \vec{U}[f] \; , \tag{2.16}$$

where $A[f] \in \mathbb{C}^{M \times N}$ is the complex-valued DFT transformed representation of the FIR matrix A. From the previous equations it follows that we can define the mixing and unmixing matrices $W[f] \equiv A^{\dagger}[f] \in \mathbb{C}^{N \times M}$. It can be shown from this that the ambiguities of the instantaneous linear BSS model that we discussed in the previous section now get transferred to the frequency domain for convolutional mixtures, so we can rewrite (2.7) as in [25] for convolutional BSS:

$$\hat{S}[f] = P[f]\Lambda[f]\hat{W}[f]\vec{X}[f], \tag{2.17}$$

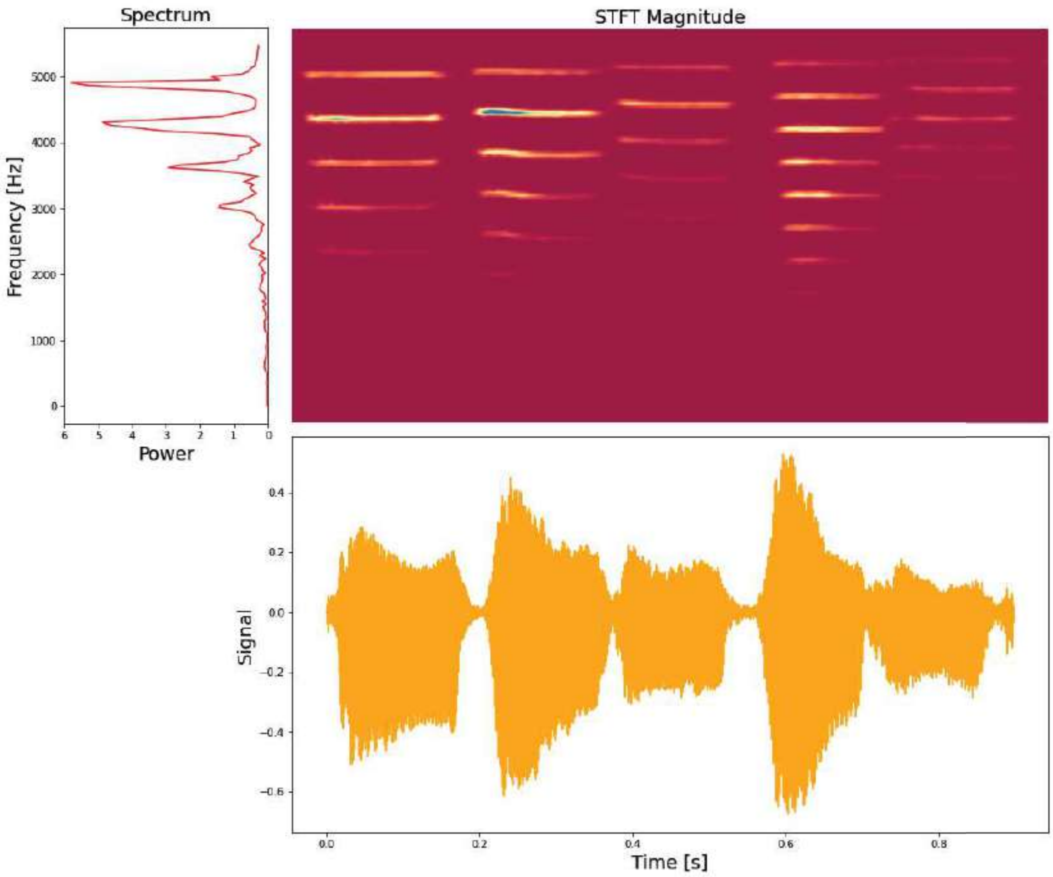


Figure 2.4: Representation of a signal in the time, frequency, and time-frequency domains

where $\Lambda[f] \in \mathbb{R}^{N \times N}$ is a diagonal scaling matrix of filters in the frequency domain, and $P[f] \in \{0, 1\}^{N \times N}$ is a permutation matrix. Supplementary to the above, a transformation of the convolutional system to the z-domain may also be derived [22].

2.3 Source Separation in the Time-Frequency Domain

Although several speech separation / enhancement frameworks exist which directly process the mixture audio signal x_{mix} in the time domain, the methodology for our separation pipeline aligns with works opting to process the signal in the time-frequency domain. More specifically, we pre-process the speech signal by transforming it using a Short-Time Fourier Transform (STFT). This representation is especially useful for non-stationary signals, where the frequency density varies over time. Voice signals are considered to be stationary over a short-time window estimated to be about 30-50 ms in duration [26].

An important property of STFT when considering signal mixtures is its additivity, meaning that, when a sum of signals is transformed using STFT, the resulting time-frequency representation is equal to the sum of the STFTs of each individual signal, i.e.:

$$x_{mix}[n] = s_1[n] + s_2[n] + \cdots + s_N[n] \Rightarrow$$
$$X_{mix}[n,f] = S_1[n,f] + S_2[n,f] + \ldots S_N[n,f]\ . \tag{2.18}$$

Furthermore, the formulations discussed earlier for convolutional BSS transfer to the time-frequency domain. Indeed, by transforming the equations (2.13) and (2.17) we obtain the de-convolution framework for this domain [23, 25]:

$$\vec{X}[n,f] = A[f]\vec{S}[n,f]\ , \tag{2.19}$$

and

$$\hat{S}[n,f] = P[f]\Lambda[f]W[f]\vec{X}[n,f],\ \text{where } W[f] = A[f]^{\dagger}\ , \tag{2.20}$$

where all matrices are similar to the ones discussed in the previous section.

An optimal unmixing matrix can be computed by minimization of some maximum likelihood (ML) criterion with the use of a gradient descent method, such as the log likelihood of the unmixed signals that are derived when applying this matrix to the input mixture[27]. Here, however, we focus our attention to methods of estimating an unmixing matrix based on minimum mean square error (MMSE) estimation.

In the local Gaussian modeling method for audio source separation [28, 29], each observed signal at an array of sensors, denoted in vector form $\vec{X}[n,f]$, is thought to be contributed to by the i-th source according to the image vector $\vec{Y_i}[n,f] \in \mathbb{C}^{M\times M}$, i.e.:

$$\vec{X}[n,f] = \sum_{i=1}^{N} \vec{Y_i}[n,f]\ . \tag{2.21}$$

In order for this to conform to our current BSS model, the image vector satisfies $\vec{Y_i}[n,f] = col_i(A[f]) \cdot S_i[n,f]$, where $col_i(A[f])$ denotes the i-th column of the mixture matrix.

Source signals are assumed to be uncorrelated. However, due to the relative positions of the microphones and sources, certain spacial correlations exist within the image vector. This fact, combined with the natural assumption of the stationary nature of the source signals at a short time frame, leads us to model the image vectors as complex-valued Gaussian random variables with zero mean and covariance matrices $\Sigma_i \in \mathbb{R}^{M\times M}$, for $i = 1,\ldots,N$, i.e.:

$$\vec{Y}_i[n,f] \sim \mathcal{N}(0, \Sigma_i[n,f])\ . \tag{2.22}$$

Furthermore, each of these covariance matrices can be refactored as:

$$\Sigma_i = v_i[n, f]R_i[f] \ , \tag{2.23}$$

where $v_i[n, f]$ is a scalar that contains the variational encoding information of source i at the time-frequency bin $[n, f]$, and $R_i[f] \in \mathbb{R}^{M\times M}$ models spatial correlation between the sources in an image. Note that the matrix $R[f]$ is static in time, meaning that the geometric configuration of the speakers and microphones does not change over time.

Under this assumption and that of independence of the source signals, the random observation vector $\overrightarrow{X}[n, f]$ also turns out to be a normally distributed Gaussian random variable with covariance matrix $\Sigma[n, f]$ modeled as:

$$\Sigma[n, f] = \sum_{i=1}^{N} v_i[n, f]R_i[f] \ , \tag{2.24}$$

thus, by Wiener filtering [30, 31], we can derive the MMSE reconstructed image signals as:

$$\hat{Y}_j[n, f] = v_j[n, f]R_j[n, f]\Sigma^{-1}[n, f]\vec{X}[n, f] \Rightarrow$$

$$\hat{Y}_j[n, f] = v_j[n, f]R_j[n, f]\left(\sum_{i=1}^{N} v_i[n, f]R_i[f]\right)^{-1}\vec{X}[n, f] \ , \tag{2.25}$$

as in [32, 29].

The last method reviewed might feel slightly foreign compared to the ones we discussed so far, estimating ”audio image signals” instead of the raw sources, but it bares notable similarity to some of the spectral masking techniques that we review in the following section. More specifically, the assumption of local stationarity of each speech signal, as well as of uncorrelatedness between them, yields the formulation for estimating a ”mask” of sorts that selects specific segments of the signal in the time-frequency domain while diminishing others. Furthermore, spectral masks are a natural extension of the MMSE estimation to the time-frequency domain.

2.4 Spectral Masking

Consider a mono-aural audio separation setup, where a collection of N source signals $\{s_i\}_{i=1}^{N}$ are to be estimated via a single channel device ($M = 1$). Recalling (2.18)

$$X_{mix}[n,f] = S_1[n,f] + S_2[n,f] + \cdots + S_N[n,f] \ ,$$

our goal is to extract any randomly selected source S_i from the mixture (perhaps as designated by a user). This leads to the following partitioning of the mixture signal:

$$\begin{gathered} X_{mix}[n,f] = S_i[n,f] + \overbrace{\sum_{j \neq i} S_j[n,f]}^{N[n,f]} \Rightarrow \\ X_{mix}[n,f] = S_i[n,f] + N[n,f] \ , \end{gathered} \tag{2.26}$$

where $S_i[n,f]$ is denoted as the target signal and $N[n,f]$ as the noise, or interference. We define a spectral separation mask as a filter in the time-frequency domain M_i that when applied via per-bin multiplication (denoted as $\odot$) to the mixture signal X_{mix} yields an approximate reconstruction of the targeted source signal:

$$M_i[n,f] \odot X_{mix}[n,f] \approx S_i[n,f] \ . \tag{2.27}$$

We next present two quintessential such formulations for time-frequency masks:

- Ideal Ratio Mask (IRM): Perhaps the most natural way to expand upon our previous discussion of applying Wiener filtering to a source separation task is by introducing the notion of the ideal ratio mask, sometimes referred to as a Wiener filter or Wiener-like mask. More specifically, the IR mask is defined as the optimal solution to the minimization of the following objective:

$$M_{IR}[n,f] = \underset{M_i[n,f]}{argmin} \left\{ \mathbb{E}\left[|S_i[n,f] - M_i[n,f] \odot X_{mix}[n,f]|^2 \right] \right\} \ . \tag{2.28}$$

  In other words, the IRM is a linear-MMSE estimator for each bin $[n,f]$ of the target source magnitude spectrogram, computed by assuming that the target and interference signals are zero-mean and mutually uncorrelated random variables. By solving the above quadratic optimization problem, we can derive:

$$M_{IR}[n,f] = \frac{|S_i[n,f]|^2}{|S_i[n,f]|^2 + |N[n,f]|^2} \ . \tag{2.29}$$

  In [33, 34], a similar formulation of this filter is suggested:

$$M_{IR}[n,f] = \left(\frac{|S_i[n,f]|^2}{|S_i[n,f]|^2 + |N[n,f]|^2}\right)^{\beta} \ , \tag{2.30}$$

where the β parameter is usually set to 0.5 in the spirit of a square root Wiener filter, used to optimally reconstruct power spectral density.

- Ideal Binary Mask (IBM): A limitation shared among all ratio masks is their indefiniteness at points where $N[n,f] \approx S_i[n,f] \approx 0$. In practical applications this can be partially resolved by adding a small real number ϵ to the denominator, but still inconsistent values of the IRM may result in unstable convergence during model training. Binary masks avoid this problem, providing a more stable learning objective.

  Binary masks offer a more simplistic take on noise filtering in the time-frequency domain, where specific spectral components pertaining to the target signal are passed on and segregated from the rest of the signal whose amplitude is set to zero. Binary-valued masks perform this action by setting their bins to one of two values $\in \{0,1\}$, applying a selection of mixture components per-bin as representative of the de-noised signal.

  An IBM filters the noisy signal by setting the time-frequency bins of the binary mask $M[n,f]$ to 1 when the target signal $S_i[n,f]$ dominates over the noise $N[n,f]$ in terms of magnitude i.e. $|S_i[n,f]|^2 > |N[n,f]|^2$, whereas the rest are set to 0. We can summarize this as:

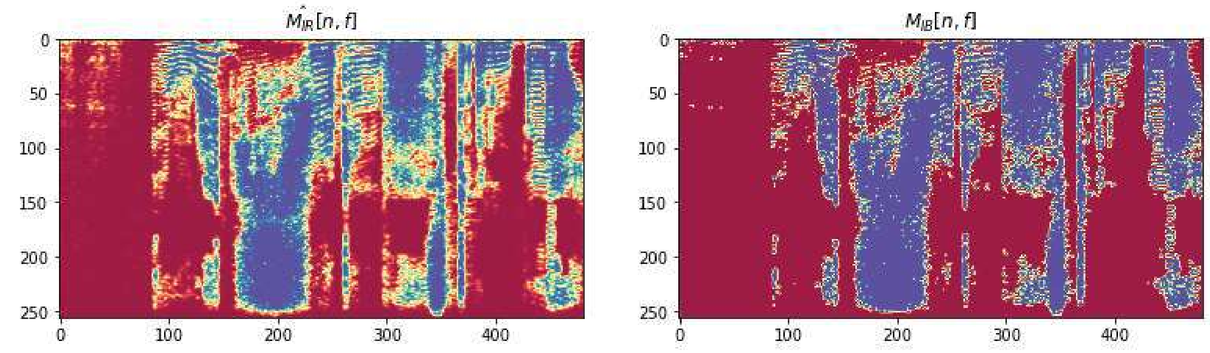
$$M_{IB}[n,f] = \begin{cases} 1, \text{ if } |S_i[n,f]|^2 > |N[n,f]|^2 \\ 0, \text{ otherwise} \end{cases} \tag{2.31}$$

  or, using a more compressed notation:

$$M_{IB}[n,f] = u(|S_i[n,f]|^2 - |N[n,f]|^2) \ , \tag{2.32}$$

  where u denotes the unit step function.

  Despite their simplicity, binary masks are popular in many different domains of signal separation applications, proving to be a robust objective target for optimal separation. Several other benefits of the IBM as outlined in [35] are their flexibility across different types of audio modalities, empirically supported psycho-acoustic correspondence, well-definiteness, and ceiling (optimal) performance. The optimality of the IBM per time-frequency bin, per time frame, and globally is shown formally by Li and Wang



(i) Ideal Ratio Mask

(ii) Ideal Binary Mask

Figure 2.5: Examples of IBM and IRM applied to a 2-speaker speech mixture

in [36] where they prove minimization of Signal-to-Noise Ratio (SNR) and MSE of the estimated signal in contrast to the ground-truth signal. In this sense, an IBM is a binary-valued Linear Minimum Mean Square Error (LMMSE) estimator.

Due to the similarity in the definitions of the IRM and IBM (both are LMMSE estimators except the IBM is binary-valued whereas the other is real-valued) they share several positive traits [37]. In fact it can be shown that both masks are special cases of the parametric Wiener filter mask [38].

However these masks are limited in their ability to invert the phase distortion induced by the mixture of audio modalities, although commonly regarded as less linguistically important for speech signals. Instead, these masks focus on isolating segments in the mixture magnitude that are representative of the target signal's magnitude. Certain spectral mask formulations such as ORM [33], PSM, [39] and cRM [40] attempt to amend this discrepancy by encoding phase information of the target and interference signals. For our model however, we limit our focus exclusively on the IRM and IBM for the sake of simplicity. Further, we will only be evaluating models trained on the IBM as an objective, as those trained with the IRM failed to yield improved results.

2.5 Spectral Masks as a Learning Objective

In this Section we present a formal description of how a spectral mask can be used as the ground-truth for estimation in a training framework of a DNN used for the task of separating select speakers from a mixture. More specifically, we formally state the goal of our architecture as a signal estimator and define an objective to minimize in order to reach its desired function.

Let us consider a set of speaker audio signals $\{s_i\}_{i=1}^{N}$, where $s_i \in \mathbb{R}^{C_s \times T_s}$, for $i = 1, \ldots, N$ consisting of T_s samples of C_s-channel audio data. For our specific application we assume a mono-aural source for each speaker i.e. $C_s = 1$, and that we have one measuring device. Let us also assume that this set is complemented by the corresponding speaker videos $v_i \in \mathbb{R}^{T_v \times C_v \times H \times W}, i = 1, \ldots, N_{spk}$ consisting of T_v consecutive $H \times W$ frames of C_v color channel camera footage. Notice that this model allows for $T_s \neq T_v$. Furthermore, it can be assumed that the audio and video modalities are only partially aligned, and in general that the number of audio frames is far greater than the number of video frames, i.e. $T_s \gg T_v$, with each video frame pertaining to a window of localized audio frames.

For the mixing model, we assume a simple linear combination model of the source signals. Revisiting (2.1), the source mixture signal can be computed as:

$$x_{mix}[n] = \sum_{i=1}^{N} a_i s_i[n] \; , \tag{2.33}$$

or equivalently in the time-frequency domain:

$$X_{mix}[n, f] = \sum_{i=1}^{N} a_i S_i[n, f] \; . \tag{2.34}$$

Following this, the problem of separating the signals generated by the several different audio sources is described as approximating a mapping $\mathcal{F}$ over the signal space [41], capable of extracting the audio modality s_i from the mixture x_{mix}:

$$x_{mix} \xrightarrow{\mathcal{F}} s_i \; .$$

We propose a generative model $\mathcal{G}$ of the separation process, capable of inferring the ideal spectral mask that separates s_i from the rest of the mixture, given audio and video modalities of the speakers:

$$\mathcal{G}\left(x_{mix}, v_i, \{v_j\}_{j \neq i}\right) \equiv \hat{M}_i[n, f] \; , \tag{2.35}$$

where $\hat{M}_i[n, f]$ is an approximate estimation of the ground-truth spectral mask defined for the separation task:

$$M_i[n,f]X_{mix}[n,f] = \hat{S}_i[n,f] \; .$$

Notice that in (2.35) the set of video modalities is partitioned to $\{v_i\} \cup \{v_j\}_{j \neq i}$, indicating that the i-th speaker video modality essentially indicates the specific speaker source that we would like to separate from the mixture. The key intuition behind incorporating the video modalities into the separation pipeline is that events in the visual domain are correlated to ones in the audio domain of the mixture, thus visual information can be used in localizing these events for each source. This essentially provides a solution to the permutation problem presented in Section 2.1 and, furthermore, allows us to infer characteristics of the audio modality from the accompanying video at each time frame, aiding in separation.

Several source separation pipelines use complex-valued time-frequency masks as the ground-truth target for unmixing the signal, however the IBM that we use in our approach is not. Therefore, our estimation of the reconstructed signal will use the possibly distorted mixture phase as an approximation of the target signal phase. This distortion could be corrected at a later stage by applying a phase correction module [6].

The reconstructed signal is derived as follows:

$$\hat{S}_i[n,f] = |M_i[n,f] \odot X_{mix}[n,f]| \angle X_{mix}[n,f] \; . \tag{2.36}$$

The estimated audio can then be extracted by applying an Inverse STFT.

The model for this estimation can be trained to minimize the bin-wise MSE between either the reconstructed signal and the reference signal, or the estimated mask and the ground-truth mask:

$$\mathcal{L}_{SPEECH}(S_i, \hat{S}_i) = \frac{1}{TF} \sum_{n,f} \Big| S_i[n,f] - \hat{S}_i[n,f] \Big|^2 \Rightarrow$$

$$\mathcal{L}_{SPEECH}(S_i, \hat{S}_i) = \frac{1}{TF} \sum_{n,f} \Big| S_i[n,f] - \hat{M}_i[n,f] \odot X_{mix}[n,f] \Big|^2 \; , \tag{2.37}$$

$$\mathcal{L}_{MASK}(M_i, \hat{M}_i) = \frac{1}{TF} \sum_{n,f} \Big| M_i[n,f] - \hat{M}_i[n,f] \Big|^2 \; , \tag{2.38}$$

respectively, where T, F are the number of time and frequency bins of the STFT. Any of the two objectives can be minimized using some gradient-based optimization algorithm such as the Adam optimizer [42].

Chapter 3

Our Proposed Approach

This chapter presents the specifics of our proposed model architecture for an audio-visual speech separator. Specifically, we discuss the separation procedure step by step, namely, the audio-visual data pre-processing for feature extraction (Section 3.1), the transformation and fusion of their corresponding latent representations as derived by an appropriate encoder for each stream (audio/visual), and finally the estimation of a spectral mask for separating the target signal from the mixture (Section 3.2).

3.1 Generating Artificial Mixtures/ Pre-processing

In this Section we outline the employed method for obtaining input sample features extracted from our datasets consisting of speaker audios and videos.

The process of generating random mixtures from a dataset D containing N_D distinct speakers labeled $\mathcal{J} = \{1, \ldots, N_D\}$ consists of first constructing an N-tuple of label-indexes $\mathcal{T} = \langle i_1, i_2, \ldots i_N \rangle$ sampled without replacement from $\mathcal{J}$. This is done in order to avoid training the model to separate a speech signal from another speech signal originating from the same speakers. We then select a random video sample v_i paired with a voice sample s_i from those available in the dataset for each speaker $i \in \mathcal{T}$.

The video modalities for each speaker $i \in \mathcal{T}$ denoted $v_i \in \mathbb{R}^{T_V \times H \times W \times C}$ may contain a different number of frames. Thus, we truncate, or pad the video data to a standard number of 60 frames, setting $T_v^{'} = 60$ which corresponds to an approximately 2-sec clip from the source video. Subsequently, these videos are cropped around the Region-of-Interest (ROI) by first locating the face area using the python ”facenet” library of a face locating multi-task CNN

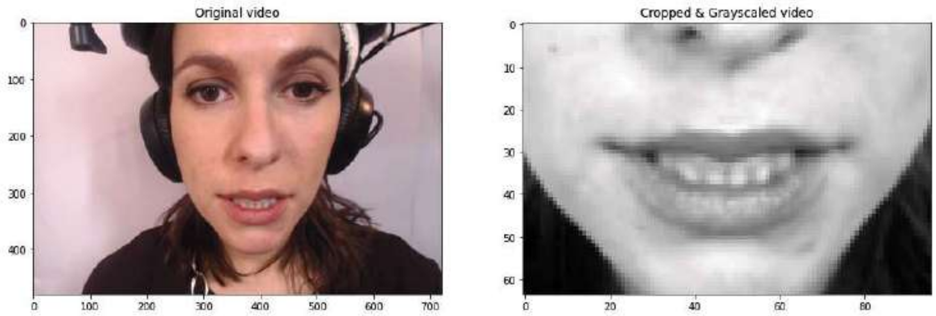
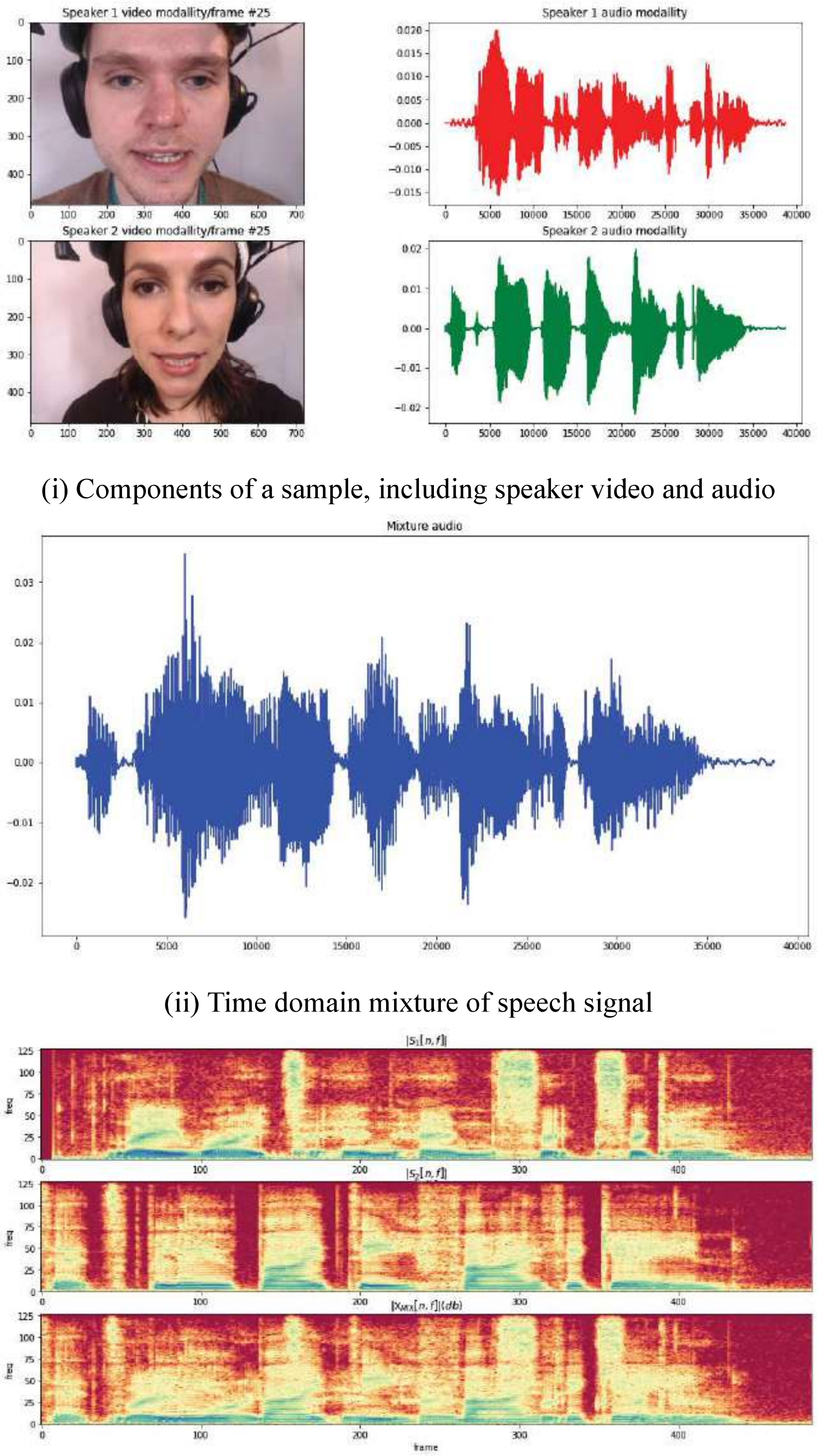


Figure 3.1: Video pre-processing

[43]. A bounding box is extracted from the first frame of the video, which is assumed to contain the speaker's face for all other frames, and a fixed 64×96 pixel sub-area is used as a boundary for the lip ROI area. Note that this simple and efficient process may induce a slightly shaky ROI especially when subjects move their heads as they speak, which may adversely affect our model's performance. Luckily, the datasets we consider in the evaluation of our model contain minimal amounts of speaker head movements. The final step in pre-processing for the visual stream is converting each frame to gray-scale. Denoting the t-th frame of the video as v_i^t, the gray-scale conversion of tri-chromatic RGB footage is performed by applying the formula $v_i^t = 0.3R(v_i^t) + 0.69G(v_i^t) + 0.11B(v_i^t)$, where $R(.), G(.), B(.)$ denote the red, green, and blue channels of the image [44]. The pre-processed video can thus be represented by a $60 \times 64 \times 96 \times 1$ tensor.

Pre-processing the audio data per speaker $s_i \in \mathbb{R}^{C \times T}$ in the mixture can be a bit tricky, since batches of audio frames need to be aligned with the video frames and different samples may contain a different ratio of audio frames to video frames. After selecting an appropriate such ratio r, a certain portion of each speech signal is cropped or padded to a length of $r \cdot 60$ audio frames that correspond to the fixed sequence of 60 video frames extracted during video pre-processing. This segment is subsequently re-sampled with a 16 kHz sampling rate. The power level of each signal is then normalized in order to resolve any scaling ambiguity: $\mathbb{E}[s_i^2] = 1$. The cropped, re-sampled, and normalized signals are finally transformed using a 512 sample STFT and a 512 sample length Hann window, which amounts to roughly 30 ms intervals in the time domain, ensuring somewhat stationary frequency distributions per STFT time frame. The stride, or hop length of the filter is set according to the available time samples so that the resulting time-frame representation has approximately 480 time frames, necessary for spoken segments between different speakers to be temporally distinguishable by the classifier. The resulting STFTs have a band of 8 KHz thought to contain most linguistic information of the speech segment. The audio modality features for each speaker after pre-processing consists of $1 \times 255 \times 48X$ tensors containing complex numbers, where X accounts



(i) Components of a sample, including speaker video and audio

(ii) Time domain mixture of speech signal

(iii) Speech signals in the time and time-frequency domains

Figure 3.2: An artificial mixture sample from the Lombard GRID dataset

for some discrepancy in the representation dependant to the audio data length of the data set.

The mixture signal is constructed in the time domain by one of the mixture models we define in Section 4.2 after re-sampling its components, and is subsequently transformed to its STFT representation by transforming by the same procedure described previously. In order to standardise the input audio data, we normalize the magnitudes of the mixture STFT to a 0-1 scale by max scaling $X_{mix} := \frac{X_{mix}}{max\{|X_{mix}|\}}$. The maximum amplitude of the spectrum is then cached for reconstruction of the separated signal at a later stage.

3.2 Architecture Description

In this Section we describe our proposed architecture and present our rationale for the design choices we have implemented.

Many architectures often employ Recurrent Neural Networks (RNN) [45] for sequence encoding from the audio and video streams, which however are slow to train and evaluate. In recent years attention mechanisms and more specifically transformer architectures employed to mitigate temporal information in sequence data have seen great success in several different machine learning applications. This motivates us to devise a simple attention fusion module capable of combining the visual and audio modalities into a heatmap-like representation of the importance of the features in the reconstruction of an audio mask and then clipping using a simple gating mechanism, to filter down the mixture audio embeddings to their most essential parts for the estimation of a spectral mask.

An outline of our architecture can be seen in Figure 3.3. Our architecture essentially consists of a multi-modal U-Net style encoder-decoder that maps aligned audio and video streams to the desired time-frequency mask. The audio stream processing pipeline is intersected in the middle part of the U-Net with the outputs from a video encoder, and modalities from the two streams are combined with the help of a fusion module capable of distinguishing similarities between the transformed video and audio, then combining the computed embeddings in an attention heatmap necessary in localizing segments of the audio embedding extracted for the mixture audio. After this, a gating mechanism clips off the least important segments of the embeddings, which are then used in constructing an approximation of the optimal time-frequency mask.

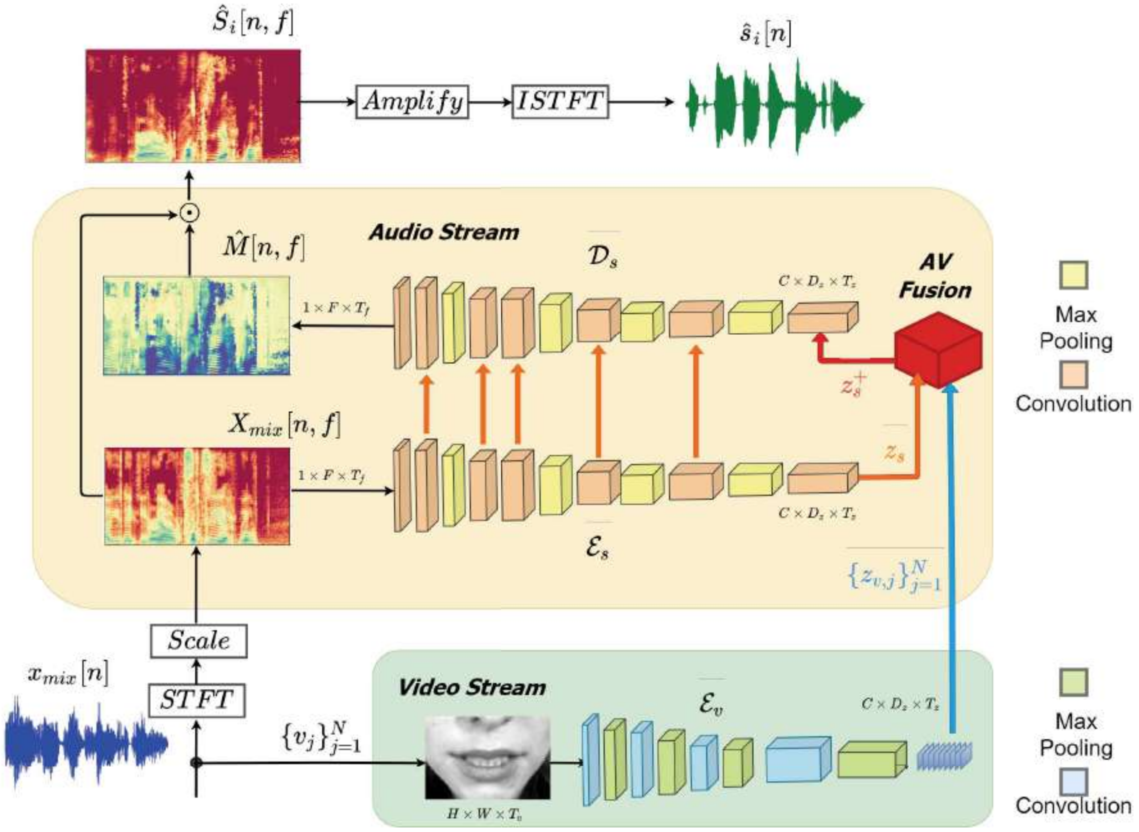


Figure 3.3: Outline of model architecture

3.2.1 Audio Stream

Throughout several speech music separation or enhancement publications, U-Net CNN architectures, which were originally conceived for the purpose of segmenting medical images [46], have been established as a tried model architecture towards separation of signals in the time-frequency domain [4, 12, 13, 14]. Furthermore, fusion of the visual modalities with the latent variables derived from the U-Net encoder has been tried before both in audio-visual speech and music signal separation with notable results [4, 12]. The main intuition behind using a U-Net for the task of separating source signals from STFT data is its inherent similarity to the problem of segmenting images: when inspecting a magnitude spectrogram of mixed audio, different speaker source signals produce visibly different patterns that may be apparently unmixable if we segment the magnitude spectrogram into different parts, assigning different segments to the speakers that they correspond to.

Our audio processing pipeline is modeled after this architecture, consisting of an encoder or ”contraction path” part $\mathcal{E}_s$ with 5 two-dimensional convolutional layers each with 3×3 kernels followed by batch normalization and ReLU activation functions. Max-pooling layers

		Conv 1	Conv 2	Conv 3	Conv 4	Conv5	Conv6	Conv Out
Encoder	Channels	8	16	32	64	128	256	300
	Pool	-	2 × 2	-	2 × 2	2 × 2	2 × 2	-
Decoder	Channels	256	128	64	32	16	8	1
	Upsample	2 × 2	2 × 2	2 × 2	_	2 × 2	-	-

Table 3.1: Audio analysis network architecture

are placed intermittently throughout the encoder part of the network. A detailed description of the encoder architecture can be found in Table 3.1. The outputs from the middle part of the U-Net are encoded audio embeddings pertaining to the time-frequency representations of the transformed features. The resulting features z_s have dimensions $C_s \times D_z \times T_z$, where $F_z < F, T_z < T$ are the compressed frequency and time dimensions of the embedding, and C_s is the number of semantically differentiated encoding channels:

$$z_s \equiv \mathcal{E}_s(X_{mix}) \in \mathbb{R}^{C_s \times F_z \times T_z}. \tag{3.1}$$

As one can observe from Table 3.1, as well as Figure 3.3, the contraction path of the network compresses the temporal-frequency dimensions of the audio features while sequentially expanding the number of channels. This process is described as ”reducing the *where* and increasing the *what*”: as resolution of the embeddings is reduced, the number of channels is increased, with each channel encoding a different aspect of the input signal. For our architecture, different numbers of channels and resolutions were tested during development of the separator, with the best model performance attained at 300 channels. As for the rest of the dimensions of the output embedding, the resolution given appropriately shaped inputs is 16×30.

Following the contraction path, lies an expansion path or the decoder of the model which inputs the audio embeddings fused with the video embeddings z_s^+ and outputs an approximate mask for separation of the target speaker's speech signal. The outputs from the decoder are passed through a sigmoid activation layer so as to have values mapped to a 0-1 range, and the final estimated separation mask is thus constructed:

$$\hat{M}(n, f) \equiv \sigma(D_s(z_s^+)) \in \mathbb{R}^{F \times T} . \tag{3.2}$$

Being that our architecture is an end-to-end model, the mask has an output equal in dimension to the input audio modality. Applying the mask to the input mixture yields the approximate representation of the signal in the time-frequency domain. The expansion path is symmetrical to the contraction path, consisting of 5 transposed convolutional layers with 3×3 kernels, batch normalization, ReLU activation, and up-sampling layers. A key technique used in constructing the spectral mask in this application is the use of skip connections by the U-Net. The outputs of each layer in the encoder are fed to their mirrored equivalents in the expansion path as an input. This makes up for the information loss induced by compression of the input images in the contraction path. By caching the outputs of each layer pre-compression, we can potentially construct a more fine-grain mask in the output.

3.2.2 Visual Stream

As stated before, the visual stream is necessary for extracting speaker-specific information that helps us assign specific extracted audio segments to the right speaker, i.e. helps us deal with the permutation problem in BSS. Furthermore, information about lip movements contains pertinent temporal information to localize specific word utterances for each speaker, as well as the overlap of utterances among different speakers. Earlier methods in AVSS pipelines have used the notion of a Visual Voice Activity Detector (V-VAD) [47] to distinguish speech activity per video frame based on local mouth movements. More complex formulations have attempted to encode cross-modal linguistic information from the visual to the audio domain into visual masks used to enhance separation [48].

	Conv1	Conv2	Conv3	Conv4	Conv5
Kernel Size	$3 \times 3 \times 3$	$3 \times 3 \times 3$	$3 \times 3 \times 3$	$3 \times 3 \times 3$	3×3
Channels	16	32	64	64	300
Pool	$4 \times 6 \times 2$	$4 \times 4 \times 1$	$2 \times 2 \times 1$	$2 \times 2 \times 1$	-

Table 3.2: Visual analysis network architecture

Here we are interested in producing embeddings from video data that match the time-frequency resolution of the audio embeddings $D_z \times T_z$. The motivation for this is to concatenate the video and audio embeddings channel-wise in order to produce the fused modalities, allowing us to use the video modality in an attention-like manner to guide the separation procedure.

Our proposed architecture consists of four 3D convolutional layers with a kernel size of $3\times3\times3$ followed by ReLU and appropriately sized max-pooling layers. The designated utility of these 4 convolutional layers is feature extraction from temporally and locally adjacent pixels. As can be seen from Table 3.2, the output channels from this part of the network is $64 \cdot D_f$ (1024 when $D_f = 16$). This practically means that each one of the D_f bins of the transformed frequency dimension of the inputs has 64 semantically different channels in the previous-to-last layer of the visual convnet. The output from the 3D convnet is $z_v^{'} \in \mathbb{R}^{64\times D_z\times 1\times T_f}$, which is vectorized or ”flattened” as $z_v^{''} = Vec(z_v^{'})$, so as to have dimension $64 \times D_z \times T_f$. The final layer is passed through a 2D convolutional layer with a 3×3 kernel, which outputs the final embedding $z_v = Conv2D(z_v^{''})$. In total, the video encoder has an output of dimension $C_v \times D_z \times T_z$. The output for the j-th speaker can be denoted as

$$z_{v,j} \equiv \mathcal{E}_v(v_j) \in \mathbb{R}^{C_v\times D_z\times T_z} \ , \tag{3.3}$$

where $\mathcal{E}_v(.)$ denotes application of the video encoder to the video modality.

In our pipeline the different speaker visual streams will use the same encoder as opposed to a dedicated one for each speaker as is used by Gabay et al. [7].

3.2.3 Fusion Module

The mechanism we use for fusing the output embeddings from both encoders is an offshoot of the Squeeze-Excitation Fusion module [11], recently used for audio and visual modality fusion for speech enhancement [10] followed by a simple gating mechanism.

For the visual modality our methodology separates the video embeddings of interfering speakers to the targeted one, which we suppose to have index i, and aggregates them one representative embedding by addition:

$$z_{v,r} = \sum_{j\neq i} z_{v,j} \ ., \tag{3.4}$$

Our intuition for this choice, as opposed to using all other speakers' visual embeddings in the fusion module, is that since the outputs of the visual stream are encoded by the same network, and their ordering should not matter (the only distinction made amongst the visual modalities is distinguishing the target speaker's video), any network that combines them should have roughly the same weights for each embedding in this set. As such, combining

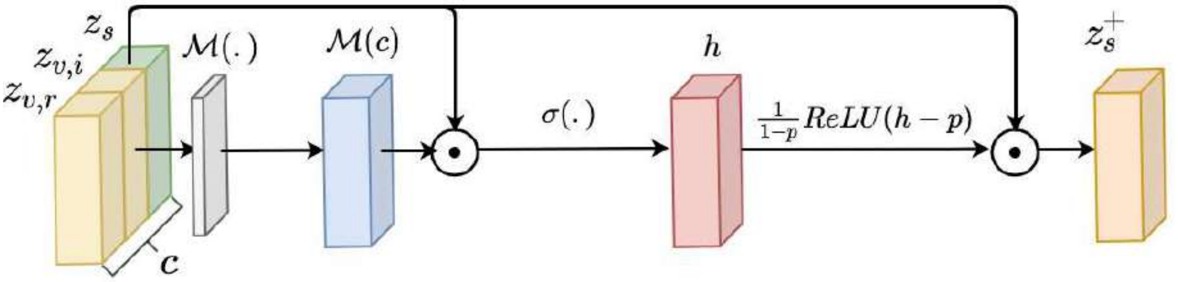


Figure 3.4: Audio-visual fusion module

them additively should produce an output not far too different than that of processing each visual modality separately, given the additivity property of convolutional networks, when their outputs are not passed through a non-linear activation function.

We denote the combined interfering speaker visual embeddings as $z_{v,r}$ as opposed to the target speaker $z_{v,i}$. Normally we would have to ”squeeze” the total resolution of the embeddings into one value per channel by averaging the outputs, however since for our architecture the resolution of the audio embeddings matches that of the video embeddings, the embeddings are simply concatenated channel-wise to produce a $(2C_v + C_s) \times D_z \times T_z$ tensor c. This way we preserve some spatio-temporal information of the embeddings. We then map the concatenated embeddings to an attention heatmap via a network consisting of a single convolutional layer $\mathcal{M}$ that takes c as an input and maps it to a $C_s \times D_z \times T_z$ embedding. This part of the network is something akin to the excitation part of the mechanism. This mapping is then point-wise multiplied to the audio embedding, and the product is passed through a sigmoid layer to produce an attention heatmap of the components of the mixture:

$$c = z_{v,i} \oplus z_{v,r} \oplus z_s \ ,$$

$$h = \sigma\left(\mathcal{M}(c) \odot z_s\right) \ . \tag{3.5}$$

This representation discriminates against or advocates for the inclusion of certain segments of the output in the fused embedding z_s^+. This embedding is produced by clipping all values of the audio embedding that are valued less than a certain threshold in the resulting heatmap and scaling the rest accordingly. This can be performed by the following gating mechanism:

$$z_s^+ = \frac{1}{1-p} \cdot ReLU\left(h - p\right) \odot z_s \ . \tag{3.6}$$

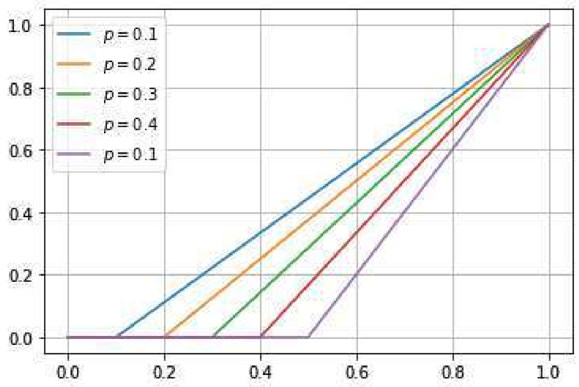
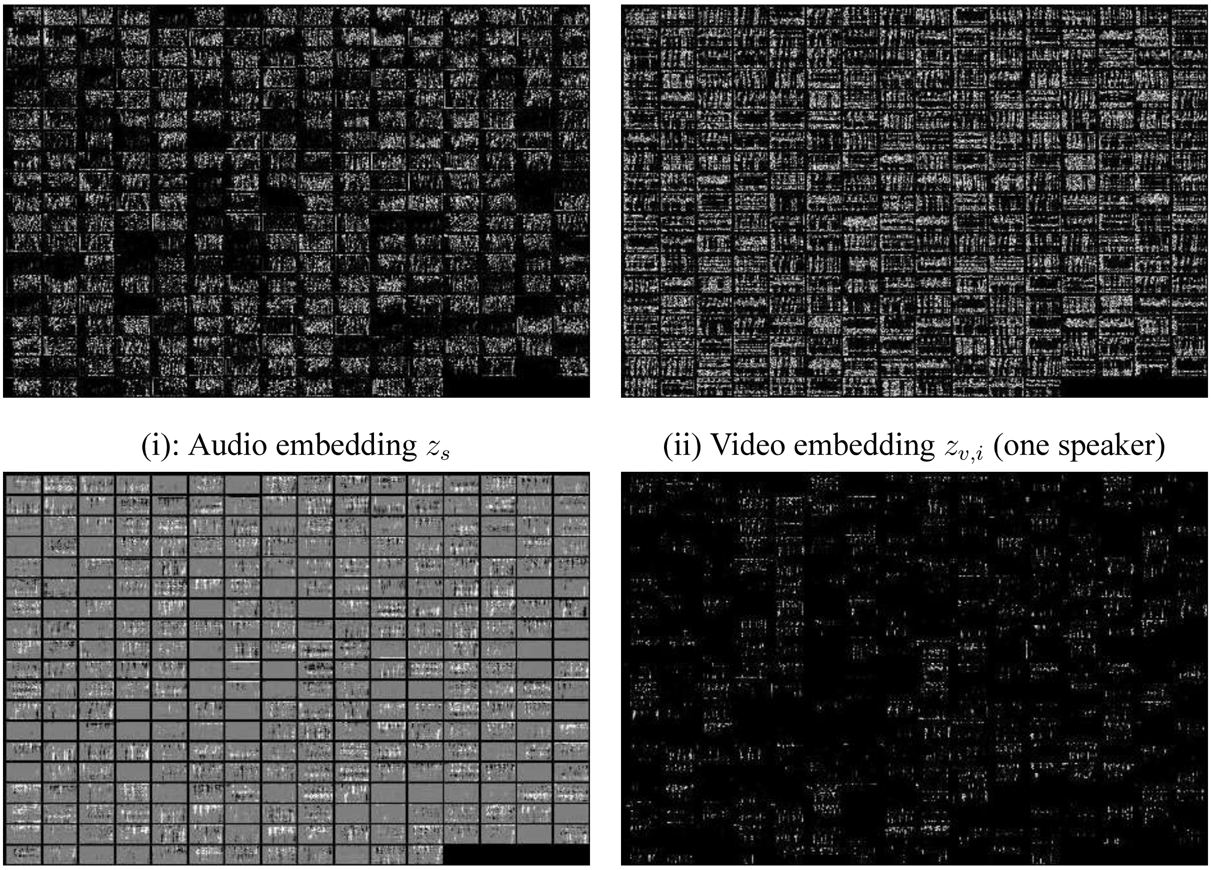


Figure 3.5: The ReLU clipping function at different threshold p values

A graph of the ReLU clipping mechanism used here, defined for different threshold values, is shown in Figure 3.5 . This works to filter out parts of the embedding that are deemed to be under a specific threshold of importance p, as the ReLU activation function is zero at points where the attention map is less than p.

The resulting clipped embeddings are fed to the decoder that constructs an appropriate mask for separating the target signal, as outlined in our description of the audio stream. To better understand the utility of each sub-network part, we present an image grid consisting of the output modalities for each channel of the corresponding encoder in Figure 3.6. Each picture shows C_s channels for a type of embedding, consisting of rectangles with dimensions $D_z \times T_z$ (Height × Width).



(i): Audio embedding z_s

(ii) Video embedding $z_{v,i}$ (one speaker)

(iii): Attention heatmap h

(iv): Fused embedding z_s^+

Figure 3.6: Channels of embeddings derived from audio-visual data

Chapter 4

Evaluation of the Proposed Model

Following the explanation of our model's architecture we now present the training procedure (Section 4.1) and experimental frameworks (Section 4.2) for our designed architecture's training and testing. Further, we present the corresponding evaluation metrics we used in the evaluation (Section 4.3), and finally we conduct a quantitative-comparative analysis of our results obtained when applying our model under different experimental premises and test datasets (Section 4.4).

4.1 Model Training and Datasets

Using artificially generated mixtures produced by the process described in the previous chapter, each model is trained on a dataset of 10000 random speaker pairings for 5 epochs. The optimizer employed is an Adam optimizer with a learning rate of 0.01 in order to minimize the joint mask-speaker loss objective which is a joint version of the objectives (2.38) and (2.37) i.e.:

$$\mathcal{L}(M_i, \hat{M}_i, S_i, \hat{S}_i) = \mathcal{L}_{MASK}(M, \hat{M}_i) + a \cdot \mathcal{L}_{SPEECH}(S, \hat{S}_i) =$$

$$\frac{1}{TF}\left(\sum_{n,f}\left|M_i[n,f]-\hat{M}_i[n,f]\right|^2 + a\cdot\sum_{n,f}\left|S_i[n,f]-\hat{M}_i[n,f]\odot X_{mix}[n,f]\right|^2\right) . \quad (4.1)$$

The datasets we train our models on are the TCD-TIMIT dataset [49] and the Lombard GRID dataset [50].

The Lombard GRID dataset is an audio-visual speech corpus of 54 individuals with both male and female vocalizations and 100 utterances per speaker. The Lombard GRID dataset provides two different viewing angles for the video modality, however we only experiment on front-facing footage. Its speech samples consist of sentences following a specific grammar from a very small vocabulary. For example, a sentence from GRID might be something like ”place red in R 4 now”. When training on the Lombard GRID, we split the dataset to 44 training speakers and 10 test speakers, roughly even in terms of their ratio of male to female vocalizing speakers.

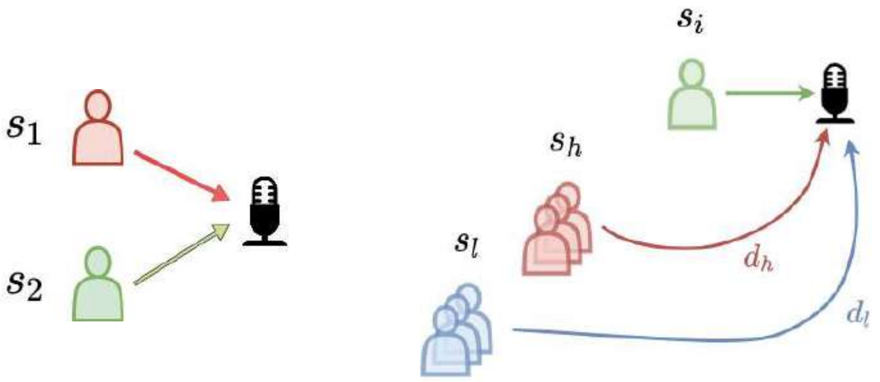
The TCD-TIMIT dataset contains speech videos and audios of up to 63 different speakers, both volunteers and trained lipspeakers (people who are trained to exaggerate mouth movements during speech in order to be easy to lipread), filmed from two different angles. Again we only use the front-facing footage. TCD-TIMIT is more linguistically rich than the GRID dataset, containing a total of 6913 sentences. When training with the TCD-TIMIT dataset, we split the dataset to 32 training speakers and 6 testing speakers with an even ratio of male to female vocalizations. Trained lipspeaker data are omitted from the training set of our model.

4.2 Experimental Setup

In this section we describe the conditions for which our model will be trained, each defined by a unique process under which the audio mixture is created, modeled after realistic conditions by which multi-speaker audio mixtures may be produced.

Broadly our experiments can be classified into two mixture scenarios:

- Scenario 1: In this scenario, two speakers are located very close to each other, with



(i) Scenario 1 (ii) Scenario 2

Figure 4.1: Speaker mixture scenarios

both of their faces visible in frontal views, speaking through the same medium, as for example in a news broadcast, where the guests-host speak over each other, resulting to:

$$x_{mix}[n] = s_1[n] + s_2[n] \ . \tag{4.2}$$

The main challenge with this setup is that the speaker audios have the same power level, which means that they are often difficult to discern from one another and inclusion of segments in the final separation mixture might occur (interference error).

This separation scenario is highly challenging, and a test to the strength of source separation architectures.

- <u>Scenario 2</u>: In this scenario, interfering speaker audios are scaled down compared to the target speaker's audio, reflecting on their non preferential placement in the sound sampling environment. An example of such a scenario is a video conference, when someone close to the microphone is having another unrelated conversation, kids are playing close by, or a TV/radio is on in the background. In such a scenario:

$$x_{mix}[n] = s_i[n] + \sum_{j \neq i} d_j s_j[n] \ , \tag{4.3}$$

where the sound sources for all the speakers except the targeted one are diminished in amplitude by some scalar random variable d_j that is distributed uniformly over an interval within $[0, 1]$. We consider two different power diminishing ranges as uniform distributions by which the d_j parameter is sampled for each speaker: A low range of scaling noise, indicating significant distance of the interfering speaker from the microphone $d_j \sim \mathcal{U}(0.3, 0.5)$, and a high range, indicating a speaker closer to the microphone, but still further away than the targeted speaker $d_j \sim \mathcal{U}(0.5, 0.8)$.

In this scenario, we will also be testing our separator under conditions that interfering speakers do not have their faces appearing in the video stream, and only the target speaker's mouth is visible.

The derived mixtures from both experimental scenarios will be passed through a designated separator model trained on data generated specifically for this task, i.e. via the same mixing process, and correspondingly the target signal will be separated by the mask produced.

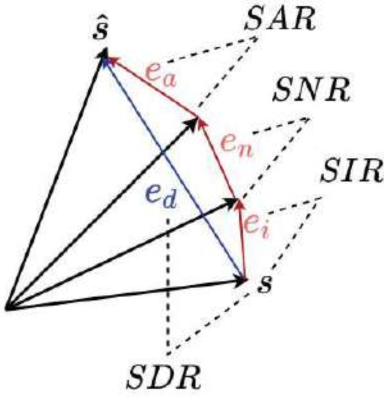


Figure 4.2: Schematic depicting relationship between SIR, SNR, SAR, and SDR

We repeat this procedure for the IBM and an oracle IRM in order to establish a theoretical upper bound for the separation task.

4.3 Evaluation Metrics

In this section we present the evaluation metrics used in assessing the reconstruction quality of the separated speaker signals derived. The quality measures considered here can broadly be classified as (i) distortion measures and (ii) perceptual quality measures.

4.3.1 Distortion Measures

Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and the Signal-to-Artifact Ratio (SAR) are commonly employed measures of noise suppression, often employed to evaluate source separation models [51]. The formulation of these three performance measures is closely related to that of the traditional SNR, considering that the total error between the ground-truth signal s and the estimated $\hat{s}$ can be broken down to 3 components:

$$\hat{s} = s + e_n + e_i + e_a \ , \tag{4.4}$$

where:

- e_n is the noise error component, denoting error that has been passed on to the reconstructed signal by inclusion of environmental noise,
- e_i is the interference error component, denoting error that has been passed on to the reconstructed signal by inclusion of interfering speakers in the mixture, and

- e_a is the artifact error component, denoting error that is attributed to erroneous distortions of the signal mixture caused by the reconstruction process.

The total sum of the error components:

$$e_d = e_n + e_i + e_a \ , \tag{4.5}$$

is called the distortion error. A pictorial description of the relation of these errors, the estimated, and the reconstructed signal is shown in Figure 4.2.

By taking the log-scale power ratio of these components compared to the target signal, we can derive meaningful quantitative metrics for the type of reconstruction error of our model, measured in dB. More specifically we define:

- Source-to-Interference Ratio (SIR):

$$SIR(s, \hat{s}) = 10 \log_{10} \left(\frac{||s||^2}{||e_i||^2} \right) \ , \tag{4.6}$$

- Source-to-Noise Ratio (SNR):

$$SNR(s, \hat{s}) = 10 \log_{10} \left(\frac{||s + e_i||^2}{||e_n||^2} \right) \ , \tag{4.7}$$

- Source-to-Artifact Ratio (SAR):

$$SAR(s, \hat{s}) = 10 \log_{10} \left(\frac{||s + e_i + e_n||^2}{||e_a||^2} \right) \ , \tag{4.8}$$

- Source-to-Distortion Ratio (SDR):

$$SDR(s, \hat{s}) = 10 \log_{10} \left(\frac{||s||^2}{||e_n + e_i + e_a||^2} \right) \ . \tag{4.9}$$

Our model will be evaluated on the SDR and SAR values attained by the separated target source approximation, as well as the SDR improvement (SDRi) attained by comparing the SDR of the separated source to that of the mixture. More specifically we define SDRi as:

$$SDRi(s, \hat{s}, x_{mix}) = SDR(s, \hat{s}) - SDR(s, x_{mix}) \ , \tag{4.10}$$

where x_{mix} is the speech signal mixture from which s was separated.

In order to calculate these measures during testing, we use the corresponding utility of the python museval library [52].

4.3.2 Perceptual Quality Measures

The perceptual quality measures considered for evaluation of our model commonly found in speech separation and enhancement literature are Perceptual Evaluation Speech Quality (PESQ) [53], Short-Time Objective Intelligibility (STOI) [54] and the Virtual Speech Quality Objective Listener (ViSQOL) [55]. Each of these defines a different process by which speech signals are mapped to a perceptual quality scale.

ViSQOL and PESQ scores are mapped on a Mean Opinion Score - Listening Quality Objective between 1 and 5, where 1 is the worst possible score and 5 the best. Note that PESQ and ViSQOL scores are shown to be highly correlated. STOI scores are percentage scores between 0 and 1, where 1 is the optimal score. The means by which we estimated each score are listed:

- For PESQ, we used an already implemented routine available at the ITU's official website, using the option +8000 for our audio data sampled at 16 kHz.

- For ViSQOL, we used the open source utility available at github using the default similarity to quality model and the option –use_speech_mode for the wide-band 8000 Hz evaluation procedure.

- For STOI, we used the python module pystoi available on github.

4.4 Results

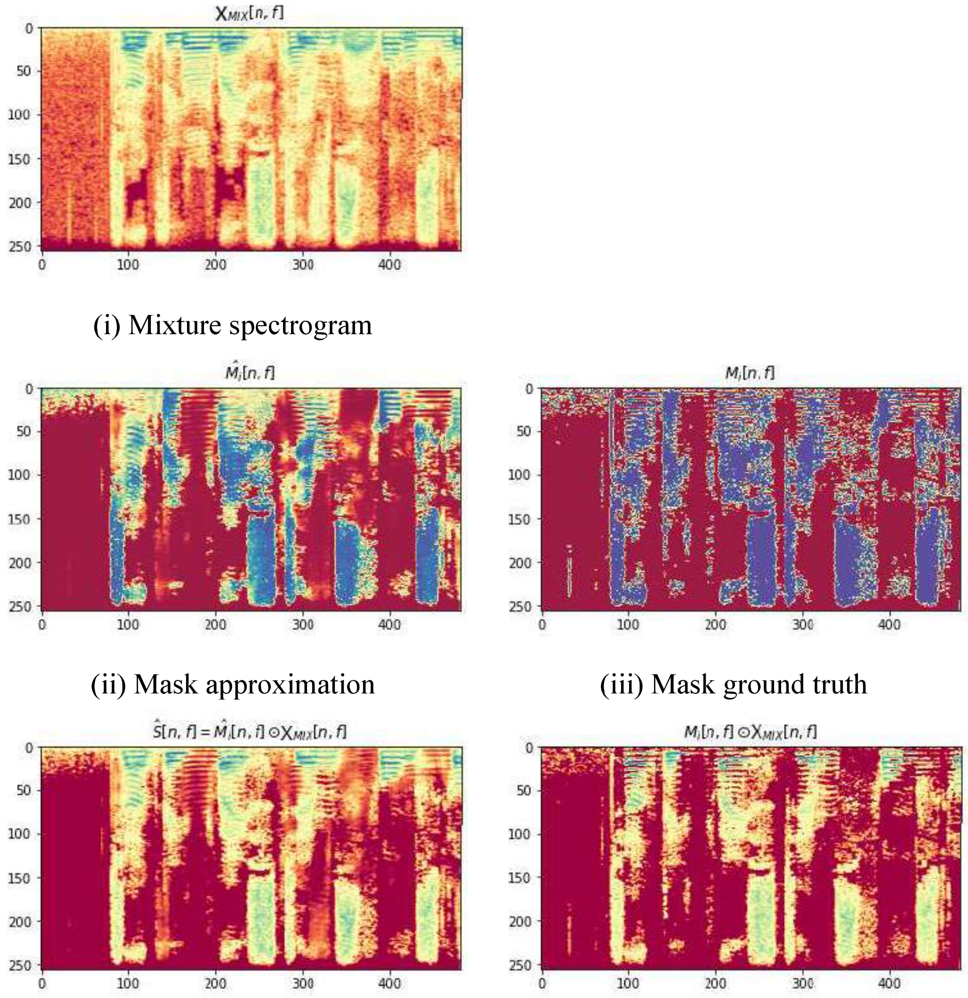
In this section we present results from different tests run for our model and contrast them with similar work done in the past by architectures designed for tasks similar to the ones we have examined.

In Table 4.1 we present a numbered list of the experiments we tested our model under, where the column ”Speakers” refers to the number of speakers in the mixture and the column ”Noise level” refers to the scaling factor sampling distributions used for the interfering speakers in scenario 2 while for scenario 1 we simply note ”Equal”. The final row ”Video” refers to the number of speaker videos available for the speakers in the mixture, with ”All” referring to the case where all speakers have their lips appear in the video feed whereas ”Target” to the case where only the target speaker appears in the video.

	Speakers	Noise Level	Videos
1	2	Low	All
2	2	High	All
3	2	Equal	All
4	2	Low	Target
5	2	High	Target
6	3	High	All
7	3	Low	All
8	3	Low	Target
9	4	Low	All

Table 4.1: List of experimental settings

In Figure 4.3 we showcase the outputs of the separator for experimental case 3, where two speakers from the TCD-TIMIT database with equal sound amplitude levels are mixed. Despite some noise passing through, notice that the separated output matches that of the IBM output, which is closely matched by its ”soft” estimation.



(i) Mixture spectrogram

(ii) Mask approximation

(iii) Mask ground truth

(iv) Separated source

(v) Ideal separation

Figure 4.3: Example of model output

A comparative breakdown of the average performance scores as well as their standard deviation is shown in Table 4.2 for the Lombard GRID corpus and in Table 4.3 for the TCD-TIMIT corpus. In general our model yields significant SDR improvement in the separated output from that of the mixture in all cases.

In low interference noise cases (experiments 1,4,7,9) the model showed admirable improvement of SDR on both datasets compared to the one feasibly achievable by a binary mask. Perceptual objectives also attained fairly high results especially for the simplest case of 2-speaker separation. For Lombard GRID, PESQ tested lower than ViSQOL scores across all cases, whereas the opposite was true for the TCD-TIMIT dataset. A notable case in SDR gain is that of 4-speaker mixtures, where the very audibly disruptive hubbub from many superimposed speakers is cleaned leading to significant gain in quality. Furthermore, our investigation in the effect of inclusion of interfering speaker videos on the GRID dataset showed that for very low noise induced by only one interfering speaker the inclusion of the video modality yields no significant improvement. For higher noise levels the 2-speaker separation task became slightly de-stabilized and yielded somewhat lower SDRi scores, as well as perceptual scores. A slightly similar result was induced in 3-speaker mixtures indicating that we may omit extraneous speaker modalities without significant loss in some cases.

	Model						Ideal Binary Mask	Ideal Ratio Mask	Mixture
	SDRi	SDR	SAR	STOI	PESQ	ViSQOL	SDRi	SDRi	SDR
1	3.23 ± 1.38	11.01 ± 1.73	12.5 ± 2.11	0.94 ± 0.03	3.07 ± 0.46	3.57 ± 0.38	4.91 ± 1.56	5.13 ± 1.57	7.78 ± 1.25
2	4.75 ± 1.98	8.35 ± 2.17	9.13 ± 2.74	0.89 ± 0.06	2.58 ± 0.53	3.00 ± 0.39	7.70 ± 1.48	7.91 ± 1.47	3.60 ± 1.17
3	6.04 ± 2.52	5.87 ± 2.52	5.82 ± 3.62	0.82 ± 0.08	2.19 ± 0.5	2.61 ± 0.42	10.2 ± 1.35	10.4 ± 1.33	-0.17 ± 0.08
4	3.24 ± 1.53	11.01 ± 1.64	12.7 ± 1.98	0.94 ± 0.03	3.12 ± 0.44	3.45 ± 0.34	4.74 ± 1.61	4.94 ± 1.63	7.76 ± 1.24
5	4.45 ± 2.18	8.05 ± 2.23	8.70 ± 3.08	0.88 ± 0.07	2.53 ± 0.55	2.95 ± 0.41	7.74 ± 1.57	7.93 ± 1.56	3.60 ± 1.22
6	5.18 ± 1.66	5.80 ± 1.75	5.86 ± 2.45	0.80 ± 0.08	2.05 ± 0.42	2.37 ± 0.33	8.88 ± 1.15	9.12 ± 1.13	0.61 ± 0.82
7	4.29 ± 1.33	8.95 ± 1.40	9.92 ± 1.76	0.89 ± 0.05	2.63 ± 0.46	2.93 ± 0.35	6.38 ± 1.25	6.61 ± 1.26	4.66 ± 0.89
8	4.14 ± 1.55	8.84 ± 1.61	9.72 ± 1.95	0.88 ± 0.06	2.50 ± 0.47	2.93 ± 0.35	6.49 ± 1.33	6.73 ± 1.34	4.71 ± 0.91
9	4.75 ± 1.17	7.74 ± 1.22	8.41 ± 1.56	0.85 ± 0.06	2.34 ± 0.45	2.60 ± 0.29	7.16 ± 1.01	7.39 ± 1.02	2.99 ± 0.72

Table 4.2: Experimental results on Lombard GRID corpus

	Model						Ideal Binary Mask	Ideal Ratio Mask	Mix
	SDRi	SDR	SAR	STOI	PESQ	ViSQOL	SDRi	SDRi	SDR
1	3.42 ± 2.34	11.16 ± 2.38	13.87 ± 2.84	0.93 ± 0.05	3.05 ± 0.47	3.03 ± 0.51	5.33 ± 2.51	5.45 ± 2.53	7.74 ± 1.24
2	5.54 ± 2.69	9.24 ± 2.64	11.04 ± 3.36	0.89 ± 0.07	2.89 ± 0.47	2.72 ± 0.46	8.57 ± 2.29	8.71 ± 2.29	3.69 ± 1.23
3	7.56 ± 2.67	7.41 ± 2.69	8.79 ± 3.55	0.85 ± 0.08	2.56 ± 0.47	2.45 ± 0.42	11.33 ± 2.17	11.5 ± 2.17	-0.15 ± 0.07
6	5.88 ± 2.06	6.45 ± 2.09	7.28 ± 2.73	0.8 ± 0.09	2.33 ± 0.41	2.24 ± 0.36	10.08 ± 1.97	10.28 ± 2	0.57 ± 0.78
7	4.5 ± 1.98	9.23 ± 2.05	10.99 ± 2.23	0.88 ± 0.07	2.71 ± 0.41	2.6 ± 0.44	7.17 ± 2.25	7.33 ± 2.28	4.73 ± 0.89

Table 4.3: Experimental results on TCD-TIMIT corpus

As for the 2-speaker equal sound amplitude mixture, our model achieves an average SDRi of 6.04 on the Lombard GRID dataset and 7.56 on the TCD-TIMIT dataset. In general, tests

(SDRi)	1	2	3	4	5
FF	3.43	4.91	5.83	3.33	4.57
FM	3.67	4.93	6.88	3.54	5.34
MF	3.17	5.31	6.37	3.38	4.23
MM	2.58	3.69	4.85	2.69	3.39

Table 4.4: Voice gender pairings effect on separation performance on Lombard GRID

TCD-TIMIT	PESQ	SDR
Gabbay [7]	2.09	0.40
Ephrat [5]	2.42	4.10
Gao [4]	2.91	10.9

GRID	SDR	SIR	SAR	PESQ
Gabbay [7] IB Mask	1.85	8.61	4.06	1.74
Gabbay [7] IR Mask	3.06	5.86	7.9	2.42

Table 4.5: Results of similar architectures on 2-speaker separation

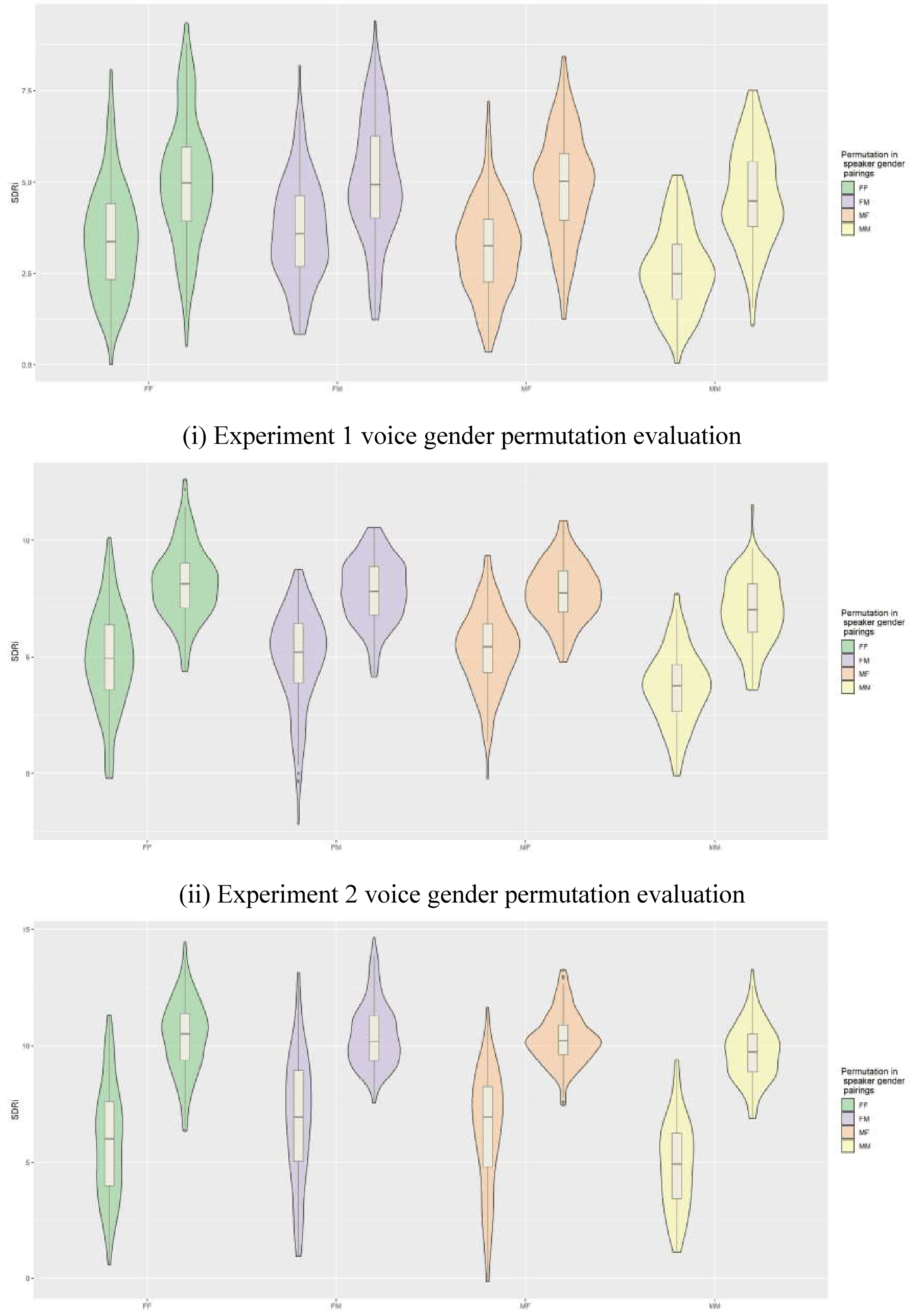
on TCD-TIMIT tend to work better for our architecture than those run on Lombard GRID across most metrics except curiously enough for ViSQOL. Notice, however, that the standard deviation in SDR improvement on this dataset is significantly higher than that on Lombard GRID. This may indicate the advantage of having a more diverse spoken vocabulary in separating speech, where it is more rare for words to coincide. Results on this experiment would put us ahead of Ephrat [5] and Gabbay's [7] work as can be seen from Table 4.5. However, our model's performance based on this test is yet to reach the state-of-the-art performance of Gao's model [4] .

An idiosyncrasy of our model is its inability to function without the use of a video modality that indicates the targeted speaker. As such, an audio-only baseline has been omitted.

Another interesting factor to investigate in 2-speaker mixtures is the effect of different gender permutations on separation performance. In [8] and [5] a comparative breakdown of model performance is presented on different kinds of speaker voice gender combinations finding that male to male voice pairings are the hardest to separate, whereas female to female voices do better, and different voice gender pairings (male to female or female to male) tend to be more easily separable .

Our model differs from the previously cited work [8], [5] in that instead of separating all speakers in the mixture our model specifically targets one speaker to separate from all the rest, meaning that beyond the gender combinations, the specific permutation i.e. which speaker is targeted matters. In Table 4.4 and in the violin plots of Figure 4.4 we have gathered results for SDR improvement across different types of such pairings for all experimental cases that involve two speakers, and note their average performance.

In general, it seems that separating female voices is easier, which seems to align with



(i) Experiment 1 voice gender permutation evaluation

(ii) Experiment 2 voice gender permutation evaluation

(iii) Experiment 3 voice gender permutation evaluation

Figure 4.4: Model performance on different permutations voice gender

the fact that female voices present a richness in frequency content greater than that of male voices making them more distinctive. Overall the female to male permutation seems to be the most favourable for separation.

As a final note on our model's performance, by inspecting the violin plots we have produced, one can clearly see that by increasing the interference magnitude against which the model has to discriminate increases variance in model performance and distances the average SDRi score from the one attained by the oracle binary mask. The score distributions of the model and the oracle are presented side by side in Figure 4.4.

Chapter 5

Conclusion

In this chapter, we draw our conclusions derived from the outcomes of the experimental evaluation of our model (Section 5.1), as well as discuss possible future improvements, experiments, and potential applications that may be tested for the architecture that we outlined (Section 5.2).

5.1 Summary

To summarize, we state some of the notable achievements as well as shortcomings of our architecture.

We managed to show that the audio-visual modality fusion method we developed can yield significant results in speech separation tasks when combined with a U-Net architecture and a 3D convolutional lip-reading Net. Another achievement of this model is its relative simplicity to the performance it yields, containing no expensive fully connected or recursive layers to model sequences, yielding good performance using only a small number of convolutional layers. Furthermore, although our network may lack in total performance compared to the current state-of-the-art, this kind of architecture could be highly scalable if adjusted to work more efficiently, with no further parameters added as speakers are added to it, with the addend that we may be able omit the video of interfering speakers and lose little accuracy, however this claim may need further testing to prove for more speakers.

Although our model succeeds on average in suppressing a significant amount of interference noise, it has a long way to go in order to meet state-of-the art standards and real world applications. Significant work that transcends the purposes of this Thesis needs to be done in order to further optimize this model until it is ready for use.

5.2 Future Work

In the following list we present some potential additions to this thesis that came up during development that we wish to see tried/implemented in future work:

- Optimizing hyper-parameters: Due to the limitations of the machine on which our model was developed and tested (a single PC with an 4Gbyte NVIDIA palit card), we skimmed the process of fine-tuning the model with the optimal hyper-parameters as training a single model requires a lot of time. These include the number of intermediate channels of the embeddings being fused, the kernel sizes of the networks, the learning objective as well as the spectral resolution of the input spectrograms and more.
- Use of a more complex mask formulation / Phase correction: Although the IR mask was originally tested as an objective for our model, it induced shakier training results and yielded worse performance across all experimental cases, despite the IR oracle mask achieving better separations across all cases. A second try at this however, perhaps with changing some training hyper-parameters may prove fruitfull. Secondly, a complex or phase sensitive mask may be used to diminish some of the adverse effects of ”phasiness” in the reconstructed audios. A different approach to correcting phase distortion could be designing a phase sub-network as in Afouras et al. [6] for correcting distortions in phase after estimating the magnitude of the signal.
- Background Noise in Separation: In the work done by Ephrat et al. [5] the separator model is evaluated in a separation scenario that can be described as joint separation / enhancement in which the model is trained in a two speaker separation scenario like (4.2) to which background noise $w[n]$ is added to the mixture, i.e.:

  $$x_{mix}[n] = s_1[n] + s_2[n] + \alpha \cdot w[n]$$

  Furthermore, in Nguyen et al. [9] a similar mixture model is assumed where a non-negative matrix factorization (NMF) model for background noise is used.

  This sort of separation under noisy conditions poses an interesting challenge for any work such as this and could be tested here as well.

- Enhancement potential: A deeper investigation into the potential of our models ability to separate speech signals from background noise signals ought to be investigated as a parallel goal for this architecture, especially considering the U-Net's tested efficiency on noise suppression tasks [14].

- Non-frontal video modality: As stated previously, the datasets we chose to evaluate our model also contain non-frontal view video recordings of the speakers. These can be used in an investigation towards view-invariant speech separation [56].

- Online source separation: A drawback of our design is its inability to be applied online, meaning that we cannot have the model operate on a live video stream, separating audio frames from background noise on-the-fly. It might be fruitful to modify our architecture to instead separate smaller sequenced chunks of audio data, instead of processing large segments entirely as is done here.

- In-the-wild data: Many in-the-wild audio-visual speech datasets exist for training and evaluating audio-visual speech separation pipelines, such as VoxCeleb [57], LRS [58], AVSpeech [5] and others, but we opted to train our model on a controlled condition setting as our initial goal. Furthering this project could include training and testing the model for datasets such as these.

Bibliography

[1] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25:975–979, 1953.

[2] S. Haykin and Z. Chen. The cocktail party problem. *Neural Computation*, 17:1875–1902, 2005.

[3] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

[4] R. Gao and K. Grauman. VisualVoice: Audio-visual speech separation with cross-modal consistency. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15495–15505, 2021.

[5] 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. *ACM Transactions on Graphics*, 37(4):1–11, 2018.

[6] T. Afouras, J. S. Chung, and A. Zisserman. The conversation: Deep audio-visual speech enhancement. In *Proceedings of Interspeech*, pages 3244–3248, 2018.

[7] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg. Seeing through noise: Visually driven speaker separation and enhancement. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3051–3055, 2018.

[8] R. Lu, Z. Duan, and C. Zhang. Audio–visual deep clustering for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1697–1712, 2019.

[9] V. N. Nguyen, M. Sadeghi, E. Ricci, and X. Alameda-Pineda. Deep variational generative models for audio-visual speech separation. In *Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2021.

[10] R. V. J. Hamid, S. Amirreza, L. I. Michael, and K. Kazuhito. MMTM: Multimodal transfer module for CNN fusion. *In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296, 2020.

[11] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[12] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba. Music gesture for visual sound separation. *Computing Research Repository (CoRR)*, abs/2004.09476, 2020.

[13] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde. Singing voice separation with deep U-Net convolutional networks. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2017.

[14] E. J. Nustede and J. Anemüller. Towards end-to-end speech enhancement with a variational U-Net architecture. *Computing Research Repository (CoRR)*, abs/2012.03594, 2020.

[15] A. Cichoki, J. Karhunen, W. Kasprzak, and R. Vigário. Neural networks for blind separation with unknown number of sources. *Neurocomputing*, 24(1):55–93, 1999.

[16] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[17] X. R. Cao and R. W. Liu. General approach to blind source separation. *IEEE Transactions on Signal Processing*, 44:562 – 571, 1996.

[18] A. Hyvarinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000.

[19] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–129, 1999.

[20] S. Jain and C. Rai. Blind source separation and ICA techniques: a review. *International Journal of Environmental Science and Technology (IJEST)*, 4:1490–1503, 2012.

[21] G. Naik and D. Kumar. An overview of independent component analysis and its applications. *Informatica*, 35:63–81, 01 2011.

[22] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra. Convolutive blind source separation methods. In J. Benesty, Y. A. Huang, and M. M. Sondhi, editors, *Springer Handbook of Speech Processing*, pages 1065–1094. Springer Press, 2008.

[23] N. Mitianoudis and M. E. Davies. Audio source separation of convolutive mixtures. *IEEE Transactions on Speech and Audio Processing*, 11(5):489–497, 2003.

[24] P. J. Smaragdis. Information theoretic approaches to source separation. Master's thesis, Masachusets Institute of Technology, May 1997.

[25] B. Rivet, S. Naqvi, and J. Chambers. Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine*, 31:125–134, 2014.

[26] P. K. Mongia and R. K. Sharma. Estimation and statistical analysis of human voice parameters to investigate the influence of psychological stress and to determine the vocal tract transfer function of an individual. *Journal of Computer Networks and Communications*, 2014:290147:1–290147:17, 2014.

[27] S. Liu, B.Wang, and L. Zhang. Blind source separation method based on neural network with bias term and maximum likelihood estimation criterion. *Sensors*, 21(3):973, 2021.

[28] M. Togami. Online speech source separation based on maximum likelihood of local gaussian modeling. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 213–216, 2011.

[29] T. T. H. Duong, N. Q. K. Duong, P. C. Nguyen, and C. Q. Nguyen. Gaussian modeling-based multichannel audio source separation exploiting generic source spectral model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):32–43, 2019.

[30] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press, 1964.

[31] J. S. Lim and A. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67:1586–1604, 1979.

[32] A. Liutkus, J. L. Durrieu, L. Daudet, and G. Richard. An overview of informed audio source separation. In *Proceedings of WIAMIS 2013 - The 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services*, pages 1–4, 2013.

[33] S. Xia, H. Li, and X. Zhang. Using optimal ratio mask as training target for supervised speech separation. In *Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 163–166, 2017.

[34] Y. Wang, A. Narayanan, and D. Wang. On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1849–1858, 2014.

[35] D. Wang. On ideal binary mask as the computational goal of auditory scene analysis. In P. Divenyi, editor, *Speech Separation by Humans and Machines*, pages 181–197. Springer, 2005.

[36] Y. Li and D. Wang. On the optimality of ideal binary time-frequency masks. In *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3501–3504, 2008.

[37] C. Hummersone, T. Stokes, and T. Brookes. On the ideal ratio mask as the goal of computational auditory scene analysis. In G. R. Naik and W. Wang, editors, *Blind Source Separation: Advances in Theory, Algorithms and Applications*, pages 349–368. Springer, 2014.

[38] R. A. Chiea, M. H. Costa, and G. Barrault. New insights on the optimality of parameterized Wiener filters for speech enhancement applications. *Speech Communication*, 109:46–54, 2019.

[39] H. Erdogan, J. Hershey, S. Watanabe, and J. Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 708–712, 04 2015.

[40] D. S. Williamson, Y. Wang, and D. Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):483–492, 2016.

[41] L. Zheng and R. Gallager. *Principles of Digital Communications I 6.450*. MIT OpenCourseWare, 2006.

[42] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Computing Research Repository (CoRR)*, abs/1412.6980, 2015.

[43] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[44] S. Chandran. Color image to grayscale image conversion. In *Proceedings of the 2010 Second International Conference on Computer Engineering and Applications*, pages 196–199, 2010.

[45] A. Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Computing Research Repository (CoRR)*, abs/1808.03314, 2018.

[46] R. Olaf, F. Philipp, and B. Thomas. U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, Wells W. M., and A. F. Frangi, editors, *Proceedings of Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer.

[47] B. Rivet, L. Girin, and C. Jutten. Visual voice activity detection as a help for speech source separation from convolutive mixtures. *Speech Communication*, 49:667–677, 2007.

[48] Q. Liu, W. Wang, P. J. B. Jackson, M. Barnard, J. Kittler, and J. Chambers. Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking. *IEEE Transactions on Signal Processing*, 61:5520–5535, 2013.

[49] N. Harte and E. Gillen. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17:603–615, 2015.

[50] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown. A corpus of audio-visual Lombard speech with frontal and profile views. *The Journal of the Acoustical Society of America*, 143, EL523, 2018.

[51] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

[52] F. R. Stöter, A. Liutkus, and N. Ito. The 2018 signal separation evaluation campaign. In Y. Deville, S. Gannot, R. Mason, M.D. Plumbley, and D. Ward, editors, *Latent Variable Analysis and Signal Separation: 14th International Conference LVA/ICA 2018, Proceedings*, pages 293–305. Springer, 2018.

[53] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2001 (ICASSP)*, volume 2, pages 749–752, 2001.

[54] C. H. Taal, R. C. Hendriks, R Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, 2010.

[55] M. Chinen, Felicia S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines. ViSQOL v3: An open source production ready objective speech and audio metric. In *Proceedings of the 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2020.

[56] A. Koumparoulis and G. Potamianos. Deep View2View mapping for view-invariant lipreading. In *Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 588–594, 2018.

[57] A. Nagrani, J. S. Chung, and A. Zisserman. VoxCeleb: A large-scale speaker identification dataset. In *Proceedings of Interspeech*, page 2616–2620, 2017.

[58] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the

wild. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3444–3453, 2017.