

TDFNet: An Efficient Audio-Visual Speech Separation Model with Top-down Fusion

Samuel Pegg^{1,†}, Kai Li^{1,†}, Xiaolin Hu^{1,2,3,*}

1. *Department of Computer Science and Technology, Institute for AI
BNRist, Tsinghua University, Beijing 100084, China*

2. *Tsinghua Laboratory of Brain and Intelligence (THBI),
IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing 100084, China*

3. *Chinese Institute for Brain Research (CIBR), Beijing 100010, China*
{peggsr10, lk21}@mails.tsinghua.edu.cn, xihu@tsinghua.edu.cn

Abstract—Audio-visual speech separation has gained significant traction in recent years due to its potential applications in various fields such as speech recognition, diarization, scene analysis and assistive technologies. Designing a lightweight audio-visual speech separation network is important for low-latency applications, but existing methods often require higher computational costs and more parameters to achieve better separation performance. In this paper, we present an audio-visual speech separation model called Top-Down-Fusion Net (TDFNet), a state-of-the-art (SOTA) model for audio-visual speech separation, which builds upon the architecture of TDANet, an audio-only speech separation method. TDANet serves as the architectural foundation for the auditory and visual networks within TDFNet, offering an efficient model with fewer parameters. On the LRS2-2Mix dataset, TDFNet achieves a performance increase of up to 10% across all performance metrics compared with the previous SOTA method CTCNet. Remarkably, these results are achieved using fewer parameters and only 28% of the multiply-accumulate operations (MACs) of CTCNet. In essence, our method presents a highly effective and efficient solution to the challenges of speech separation within the audio-visual domain, making significant strides in harnessing visual information optimally.

Index Terms—Audio-Visual, Multi-Modal, Speech-Separation

I. INTRODUCTION

Speech separation is the process of extracting distinct audio streams from an audio recording containing one or more speakers [2]. Consider a scenario where we have a microphone positioned in the center of a room capturing the voices of two individuals, A and B. These speakers may talk simultaneously or one after the other, with varying volume levels and distances from the microphone. It is crucial to account for all these factors. The objective of a speech separation model is to split the audio recording into two separate streams, each containing the audio from a single speaker. Ideally, one output stream would exclusively contain the voice of speaker A or B.

Speech separation is commonly referred to as the “cocktail party problem” [3], [13]. At social gatherings, such as cocktail parties, our natural inclination is to concentrate on a particular individual while filtering out the surrounding conversations. Over the past decade, High quality speech separation has



Fig. 1. Audio-visual speech separation process. From top to bottom: video frames and mixed inputs, cutting out lip regions from the video, TDFNet speech separation results.

become increasingly crucial due to the widespread adoption of automated systems and voice assistants such as Apple’s Siri and home devices such as Amazon Alexa or Google Home [12], [17], [18], [23], [31], [40].

Methods based on architectures designed for modeling sequences, such as Recurrent Neural Networks (RNNs) [29], or extracting local patterns such as Convolutional Neural Networks (CNNs) [30], have proven to be adept at handling

[†] Samuel Pegg and Kai Li have contributed equally to this work.

* Corresponding author.

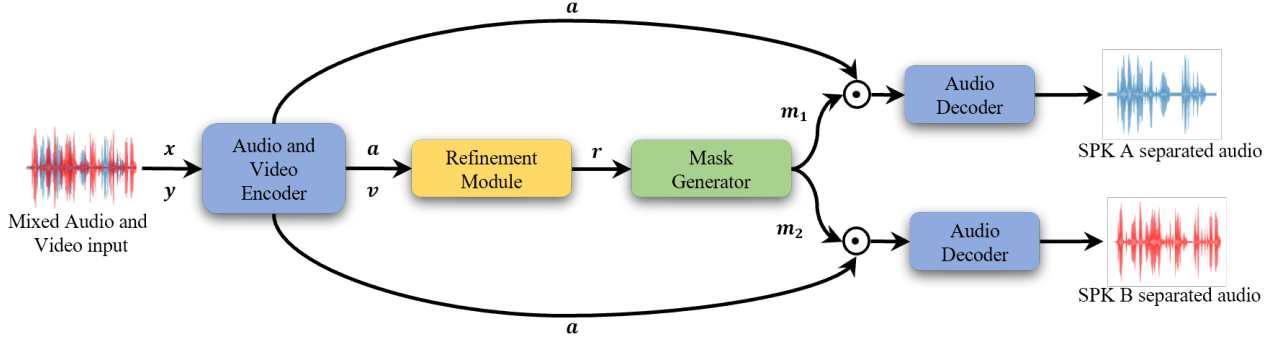


Fig. 2. TDFNet separation pipeline. The audio and video inputs \mathbf{x} and \mathbf{y} are encoded by E_a and E_v respectively to produce the feature maps \mathbf{a} and \mathbf{v} , which are sent to the refinement module R to be fused and then further processed. The mask generator M then takes these refined features \mathbf{r} and generates masks \mathbf{m}_i , which are multiplied by the encoded audio input \mathbf{a} in term to produce a separation. Finally, the decoder decodes each of the separated encoded audios. The figure above uses $n_{spk} = 2$ speakers.

the complexities in speech signals. However, methods relying on auditory signals alone, known as audio-only speech separation (AOSS) methods, have become quite saturated in recent years, with new models showing only incremental performance increases. One notable avenue to bolster the robustness of speech separation is to integrate multi-modal information [4], [41]. For humans, the integration of auditory information (speech signals) and visual stimuli (such as lip movements) fundamentally alters our perception of language [7], [34], [35].

Audio-visual speech separation (AVSS) is similar to speaker extraction [10], [50], [52], which uses one speaker’s voiceprint to create a separation. However, whereas voice extraction requires enrolling a target speaker in advance and gathering their voiceprint, visual information can be entirely captured in real-time, as shown in Figure 1. Visualvoice [9] and CTCNet [24] both use an encoder-decoder structure to improve the separation performance, but they do not fully exploit this structure. In addition, integrating audio and visual information can be computationally expensive, making them less practical for real-time applications.

In order to efficiently extract and integrate features from different modalities, we propose a novel AVSS model called **TDFNet**. It uses encoded audio and visual information as inputs for a series of TDANet-based blocks [25], called TDFNet Blocks. These blocks build a hierarchical structure where the lower levels have higher temporal resolution. Information from different temporal resolutions moves freely, and is aggregated and fused with the visual information. We evaluated TDFNet extensively on the LRS2 benchmark dataset to show that it outperforms various competing baseline models and achieves state-of-the-art performance using fewer parameters and with a lower computational cost.

II. RELATED WORK

A. Audio-Only Speech Separation

Initially, speech separation employed the time-frequency (TF) representation of the mixed audio, estimated from the waveform using the short-time Fourier transform (STFT).

Pioneering methods utilized matrix factorization [39] and heuristic techniques [6] to cluster the TF bins of each speaker. However, the performance of these models was either poor or speaker-dependent. With the development of deep learning and the introduction of permutation invariant training (PIT [53]) to solve the permutation problem, the speech separation space has become increasingly competitive. Researchers soon migrated from the TF representation of audio to a time domain only representation using a convolutional encoder, resulting in high performance methods such as DualPathRNN [29], Sepformer [44], Wavesplit [54], etc. Recently, the multi-scale speech separation model AFRCNN [16] was proposed as a balance between efficiency and separation performance. TDANet [25] adds top-down attention [32], [42] to AFRCNN’s encoder-decoder architecture to obtain state-of-the-art performance with a greatly reduced computational effort. Both of these models were inspired by the brain.

Some researchers suggest that the datasets used in the audio-only setting are often idealized, lacking reverberation, interfering background sounds or random noise. A recent study [25] found that many top models become average when faced with these more challenging datasets. Furthermore, audio-only models encounter significant challenges when confronted with scenarios involving three or more speakers [14], or when the number of speakers is unknown [10], [50], [52].

B. Audio-Visual Speech Separation

The audio-visual speech separation field combines audio-only speech separation with video data to improve performance on noisy and challenging datasets. Recent neuroscience studies have demonstrated that the human brain effectively addresses the cocktail party problem by leveraging additional visual cues from the eyes [26], [43]. Building upon this notion, some researchers [1], [8], [9], [22], [27], [33], [36], [51] have endeavored to incorporate visual information into the paradigm, aiming to enhance the quality of audio separation. These efforts have culminated in the current state-of-the-art for audio-visual speech separation, namely CTCNet

[24]. However, CTCNet, due to the complexity of its multi-scale fusion operations, introduces a substantial amount of computation which limits its applicability in real-world scenarios. To address this, we utilize TDANet’s architecture for our visual and auditory feature extraction networks, which employ progressive multi-scale fusion, effectively reducing the computational cost of the multi-scale fusion.

III. TDFNET

For a black and white 25 fps video containing n_{spk} speakers, the inputs for our AVSS pipeline are: a sequence of video frames $\mathbf{y} \in \mathbb{R}^{1 \times L_v \times H_{in} \times W_{in}}$ containing the lip movements of the desired speaker, and the mixed audio stream $\mathbf{x} \in \mathbb{R}^{1 \times L_a}$, where L_v , H_{in} , W_{in} and L_a denote the number of video frames, image height, width, and the length of audio, respectively. We assume that the mono-aural (single channel) audio of the video \mathbf{x} consists of linearly superimposed voices $\mathbf{s}_i \in \mathbb{R}^{1 \times L_a}$ for $i \in [1, n_{spk}]$ such that

$$\mathbf{x} = \boldsymbol{\epsilon} + \sum_{i=1}^{n_{spk}} \mathbf{s}_i, \quad (1)$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^{1 \times L_a}$ is some presence of background noise.

We propose TDFNet, which consists of five main modules: a video encoder (Section III-A), an audio encoder (Section III-B), a refinement module (Section III-C), a mask generator (Section III-D) and an audio decoder (Section III-E). The pipeline of TDFNet is briefly described below (see Figure 2).

- 1) The raw audio and visual signals are sent to respective encoders to obtain rich feature maps. We denote the audio encoder $E_a(\cdot)$ and the video encoder $E_v(\cdot)$:

$$\begin{aligned} \mathbf{a} &= E_a(\mathbf{x}), \quad \mathbf{a} \in \mathbb{R}^{C_a \times T_a}, \\ \mathbf{v} &= E_v(\mathbf{y}), \quad \mathbf{v} \in \mathbb{R}^{C_v \times T_v}, \end{aligned}$$

where the C_a and C_v denote the dimensions of visual and auditory features. Historically, large values for C_a and C_v are used as this leads to a more detailed feature representation. However, this increases the computational cost of the refinement module. In order to mitigate this cost while maintaining performance, we design a “bottleneck” by using fewer channels $B_a \leq C_a$ and $B_v \leq C_v$. This is achieved by using 1D convolutional layers $F_a(\cdot)$ and $F_v(\cdot)$ with kernel size 1. This can be written

$$\mathbf{a}' = F_a(\mathbf{a}), \quad \mathbf{v}' = F_v(\mathbf{v}).$$

- 2) The refinement module $R(\cdot, \cdot)$, takes the audio and visual information and uses the combined information to generate a refined feature map:

$$\mathbf{r} = R(\mathbf{a}', \mathbf{v}'), \quad \mathbf{r} \in \mathbb{R}^{B_a \times T_a}$$

We can see that the output dimensions match the dimensions of the *audio* input. This is because the refinement module first fuses the visual information *into* the audio information, and then further processes these combined

multimedia features. This will be explained in Section III-C.

- 3) This refined feature map is used to generate masks for each speaker. Let $M(\cdot)$ denote a mask generating function.

$$\{\mathbf{m}_1, \dots, \mathbf{m}_{n_{spk}}\} = M(\mathbf{r}),$$

where $\mathbf{m}_i \in [0, 1]^{C_a \times T_a}$ for all $i \in [1, n_{spk}]$ and $[0, 1]$ is the range of real numbers from 0 to 1. Note that each mask has the same dimensions as the encoded audio \mathbf{a} .

- 4) The encoded audio input is multiplied element-wise by each mask in turn, resulting in the separated encoded audio for each speaker,

$$\mathbf{z}_i = \mathbf{a} \odot \mathbf{m}_i, \quad \mathbf{z}_i \in \mathbb{R}^{C_a \times T_a} \quad \forall i \in [1, n_{spk}],$$

where \odot represents element-wise multiplication.

- 5) The separated audio feature map for each speaker is given to the decoder $D_a(\cdot)$. The decoder returns the separated audios for each speaker as a waveform,

$$\hat{\mathbf{s}}_i = D_a(\mathbf{z}_i), \quad \hat{\mathbf{s}}_i \in \mathbb{R}^{1 \times L_a} \quad \forall i \in [1, n_{spk}].$$

A. Video Encoder

The video encoder takes the gray-scale video frames \mathbf{y} and uses the pre-trained lip-reading model used in CTCNet [24], called CTCNet-Lip, to extract visual features. It consists of a backbone network for extracting features from the image frames and a classification sub-network for word prediction. The backbone network includes a 3D convolutional layer and a standard ResNet-18 network. The frames are convolved with P 3D-kernels with size $5 \times 5 \times 1$ and spatial stride size $2 \times 2 \times 1$ to obtain a rich feature map. Each feature map is passed through the ResNet-18 network, and the resulting feature maps are passed to the classification sub-network as words for word prediction. After lip-reading pre-training, the backbone network of the CTCNet-Lip is fixed for extracting visual features, and the classification sub-network is discarded. We encourage readers to read the original paper for more details.

B. Audio Encoder

The audio encoder E_a takes the input audio \mathbf{x} and produces a feature map using a 1D-convolution with C_a output channels and kernel size K_a , followed by global layer normalization (gLN) [30] and ReLU activation. Since a stride greater than one is used, this technically changes the T_a dimension to a smaller value, compressing the audio, but for simplicity we do not change the notation.

C. Refinement Module

The refinement module aims to take the audio and the video embeddings, fuse them together and then further process and refine the resulting multi-modal features. This way, we combine the audio and visual information in order to increase the separation performance. As is seen several times in the audio-only domain [16], [25], we take an iterative approach.

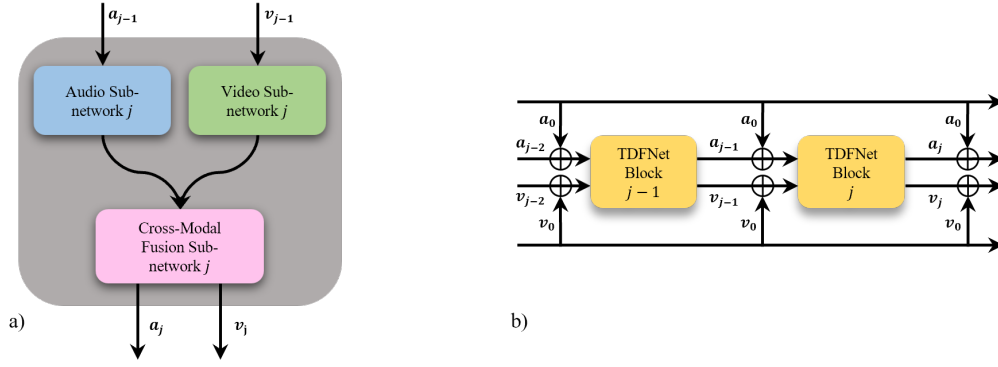


Fig. 3. The j^{th} TDFNet Block. (a) The internals of a single TDFNet block. (b) The stacked TDFNet blocks with residual connection from the first iteration.

In total, we use R_a iterations with fusion occurring R_f times, where $R_f \leq R_a$.

The refinement module consists of three networks that work together in order to produce the best possible amalgamation of data:

- α_j , the **audio sub-network** at iteration j .
- β_j , the **video sub-network** at iteration j .
- γ_j , the **cross-modal fusion sub-network** at iteration j .

These three networks do not change the dimensions of their inputs, which is key for their inter-compatibility. For the *fusion* iterations, for $j \in [1, R_f]$, one iteration consists of all three modules working together for form a TDFNet block. For the subsequent audio-only iterations, for $j \in [R_f, R_a]$, the TDFNet block will consist of *only* the audio sub-network.

Firstly, let \mathbf{a}_{j-1} and \mathbf{v}_{j-1} be the inputs for α_j and β_j , where the “0th” iteration is simply taking the outputs of the audio and video “bottleneck” convolutions:

$$\mathbf{a}_0 := \mathbf{a}' \in \mathbb{R}^{B_a \times T_a}, \quad \mathbf{v}_0 := \mathbf{v}' \in \mathbb{R}^{B_v \times T_v}. \quad (2)$$

The first iteration is the most simple. The audio features are passed to the audio sub-network and the video features are passed to the video sub-network. The outputs are then passed to the fusion module which fuses the audio information with the visual information, and the visual information with the audio information. In the proceeding layers, residual connections are added to improve training.

$$(\mathbf{a}_1, \mathbf{v}_1) = \gamma_1(\alpha_1(\mathbf{a}_0), \beta_1(\mathbf{v}_0)) \quad (3)$$

$$(\mathbf{a}_j, \mathbf{v}_j) = \gamma_j(\alpha_j(\mathbf{a}_{j-1} + \mathbf{a}_0), \beta_j(\mathbf{v}_{j-1} + \mathbf{v}_0)) \quad (4)$$

for $j \in [2, R_f]$. The interactions can be seen in Figure 3.

After R_f fusion iterations, we deem the video and audio signals to be sufficiently ‘fused’ together into the audio signal $\mathbf{a}_{R_f} \in \mathbb{R}^{B_a \times T_a}$. For subsequent iterations, the signal is continuously refined using *only* the audio sub-network. This simplifies Equation 4 to the one below.

$$\mathbf{a}_j = \alpha_j(\mathbf{a}_{j-1} + \mathbf{a}_0), \quad (5)$$

for $j \in [R_f + 1, R_a]$.

1) *Audio and Video Sub-Networks*: Both α_j and β_j are defined by the same architecture inspired by TDANet [25] and consists of three main sections:

- 1) Bottom-up Down-sampling
- 2) Recurrent Operator
- 3) Top-down Fusion

The details of these sections will be covered in Section IV. We define

$$\begin{aligned} \mathbf{X}_j &= \alpha_j(\mathbf{a}_{j-1} + \mathbf{a}_0) && \in \mathbb{R}^{B_a \times T_a}, \\ \mathbf{Y}_j &= \beta_j(\mathbf{v}_{j-1} + \mathbf{v}_0) && \in \mathbb{R}^{B_v \times T_v}, \end{aligned}$$

as the outputs of the audio and video sub-networks, and hence the two inputs for the cross-modal fusion sub-network.

2) *Cross-Modal-Fusion Sub-network*: This module is in charge of fusing the audio features into the video features, and the video features into the audio features. Let κ define a 1D-convolution with kernel size 1 followed by a gLN layer, and let ϕ denote nearest neighbor interpolation. \parallel is the concatenation operation acting along the channel dimension. Wielding these definitions, we can define the cross-modal fusion sub-network γ_j that returns two outputs, and hence we can write the full expression for the j^{th} iteration as

$$\gamma_j(\mathbf{X}_j, \mathbf{Y}_j) = (\mathbf{a}_j, \mathbf{v}_j), \quad (6)$$

where

$$\begin{aligned} \mathbf{a}_j &= \kappa(\mathbf{X}_j \parallel \phi(\mathbf{Y}_j)) && \in \mathbb{R}^{B_a \times T_a}, \\ \mathbf{v}_j &= \kappa(\mathbf{Y}_j \parallel \phi(\mathbf{X}_j)) && \in \mathbb{R}^{B_v \times T_v}. \end{aligned}$$

As we can see, the video features are interpolated to match the audio dimensions:

$$B_v \times T_v \implies B_v \times T_a.$$

The output is concatenated with the audio features and then passed through a convolution layer to take the dimensions back to the dimensions of the audio input:

$$(B_a + B_v) \times T_a \implies B_a \times T_a.$$

The video fusion result is similar.

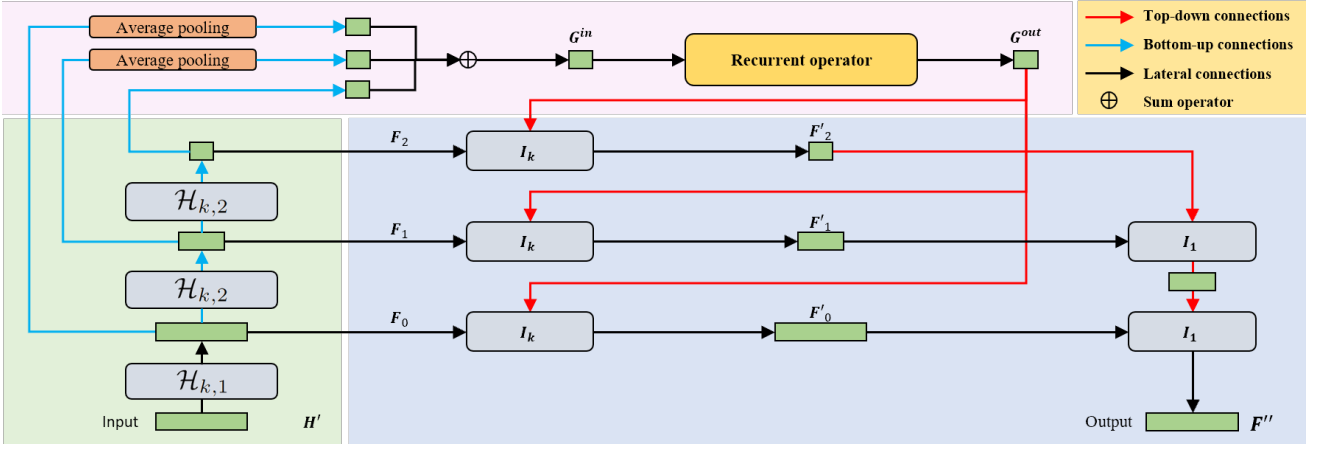


Fig. 4. The core architecture of the audio and video sub-networks. The input is either the audio or the visual features, reduced to the hidden dimension D . The bottom-up down-sampling in the green block uses consecutive convolutions with stride 2 to compress the data to increasingly small temporal resolutions. The recurrent operator in the pink block fuses the information to formulate a global perspective. The top-down fusion in the blue block combines the global information at different temporal resolutions, and then fuses them all together into a single feature map.

D. Mask Generator

The mask generator is tasked with taking the output of the refinement module \mathbf{r} and transforming it into a set of masks \mathbf{m}_i for $i \in [1, n_{spk}]$ and multiplying each mask with the encoded audio \mathbf{a} . The mask generator is characterized by a 1D-convolution that takes the channels from B_a to $n_{spk} \times C_a$. We then apply an output gate, which involves the element-wise multiplication of the $Tanh$ and sigmoid σ activations of two convolutions,

$$\begin{aligned} \mathbf{X} &= \text{ReLU}(\text{Conv}(\text{PReLU}(\mathbf{r}))), \\ \mathbf{Y} &= \text{Tanh}(\text{Conv}(\mathbf{X})) \odot \sigma(\text{Conv}(\mathbf{X})). \end{aligned} \quad (7)$$

Next, we split $\mathbf{Y} \in \mathbb{R}^{(n_{spk} \times C_a) \times T_a}$ across the channels into n_{spk} parts to get the masks $\mathbf{m}_i \in \mathbb{R}^{C_a \times T_a}$ for each speaker i . Finally, we compute

$$\mathbf{z}_i = \mathbf{a} \odot \mathbf{m}_i \quad \text{for } i \in [1, n_{spk}], \quad (8)$$

to get the separated audios of each speaker, as a feature map.

E. Decoder

The decoder mirrors the audio encoder of TDFNet in terms of stride, padding and kernel size, and hence we get the equation

$$\hat{\mathbf{s}}_i = \text{TConv}(\mathbf{z}_i) = D_a(\mathbf{z}_i) \quad \text{for } i \in [1, n_{spk}], \quad (9)$$

where TConv is now a 1D transposed convolution that converts the n_{spk} separated feature maps into n_{spk} waveform audio streams.

IV. AUDIO AND VIDEO SUB-NETWORK STRUCTURE

We have described the entire pipeline of the proposed TDFNet model without details of the audio and video sub-networks. These modules use a TDANet-like structure [25] and since the design for the audio and video sub-networks is the same, we can drop the a and v subscripts from our notation for simplicity. For more specific details of the

TDANet implementation, we refer readers to the original paper [25]. Here we will give a brief overview, focusing on our modifications.

The input to the audio and video sub-networks: $\mathbf{a}_{j-1} + \mathbf{a}_0$ or $\mathbf{v}_{j-1} + \mathbf{v}_0$, is first bound in place using a depth-wise convolution, and then converted to a lower “hidden” dimension D using another 1D convolution with kernel size 1. This will be the input for the main section of TDANet, which can be categorized into three important phases:

- 1) the bottom-up down-sampling process
- 2) the recurrent operator
- 3) the top-down fusion process

as shown in Figure 4.

Bottom-up Down-sampling: For kernel size k and stride s , we next define the normalized depth-wise convolution,

$$\mathcal{H}_{k,s}(\mathbf{X}) := \text{Norm}(\text{DWConv}_{k,s}(\mathbf{X})). \quad (10)$$

The bottom-up down-sampling process uses stacked $\mathcal{H}_{k,2}$ layers to obtain the multi-scale set \mathbf{F} with different temporal resolutions:

$$\mathbf{F} = \left\{ \mathbf{F}_0 \in \mathbb{R}^{D \times T}, \dots, \mathbf{F}_i \in \mathbb{R}^{D \times \frac{T}{2^i}}, \dots, \mathbf{F}_q \in \mathbb{R}^{D \times \frac{T}{2^q}} \right\}.$$

Recurrent Operator: To extract a global view from these features, we down-sample all elements in the set \mathbf{F} to the dimensions of the smallest element using adaptive average pooling. Next, we sum all the down-sampled features to generate the global feature map \mathbf{G}^{in} :

$$\mathbf{G}^{in} = \sum_{i=0}^q p(\mathbf{F}_i), \quad \mathbf{G}^{in} \in \mathbb{R}^{D \times \frac{T}{2^q}}. \quad (11)$$

To allow the model to understand the complex relationships between the different time steps, a recurrent operator \mathcal{R} is applied along the temporal dimension of \mathbf{G}^{in} :

$$\mathbf{G}^{out} = \mathcal{R}(\mathbf{G}^{in}), \quad \mathbf{G}^{out} \in \mathbb{R}^{D \times \frac{T}{2^q}}. \quad (12)$$

The recurrent operator refers to a sequence modelling structure such as a transformer, recurrent neural network (RNN), long-short-term-memory network (LSTM [15]) or gated recurrent unit (GRU [5]).

For the transformer (see Figure 5), we first use multi-head self-attention (MHSA) [48] followed by a drop block layer [11] and a feed forward network (FFN). The FFN consists of three stacked 1D convolutional layers with kernel sizes $\{1, k, 1\}$, and number of channels $\{D, 2D, D\}$ respectively. Residual connections are added at each stage, so for the transformer the recurrent operator \mathcal{R} is defined as:

$$\begin{aligned} \mathbf{G}^{mid} &= \text{MHSA}(\mathbf{G}^{in}) + \mathbf{G}^{in}, \\ \mathbf{G}^{out} &= \text{FFN}(\mathbf{G}^{mid}) + \mathbf{G}^{mid}. \end{aligned} \quad (13)$$

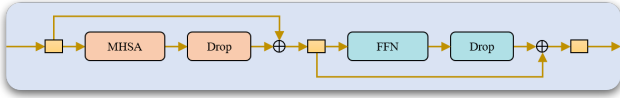


Fig. 5. Attention mechanism. This is a diagrammatic view of Equation 13.

For the other recurrent operators (RNN, LSTM and GRU) we remove the FFN and the Drop Block [11] layers, but keep the residual connection. We use bidirectional RNNs with the hidden dimension set to the input dimension D , hence we also add dropout and a linear projection layer \mathcal{P} from $2D$ channels back to D channels such that the RNN does not change the dimensions of \mathbf{G}^{out} . This alters Equation 13 to:

$$\mathbf{G}^{out} = \mathcal{P}(\text{RNN}(\mathbf{G}^{in})) + \mathbf{G}^{in}. \quad (14)$$

Top-Down Fusion: Let I_k define the Injection Sum [25] with kernel size k (see Figure 6). Then top-down fusion is defined in two steps. First, the global information \mathbf{G}^{out} is fused with each of the multi-scale local features:

$$\mathbf{F}'_i = I_k(\mathbf{F}_i, \mathbf{G}^{out}), \quad \mathbf{F}'_i \in \mathbb{R}^{D \times \frac{T}{2^i}} \quad (15)$$

for $i \in [0, q]$. Next, this multi-scale global-and-local-fusion is collapsed into a single feature map that has a broad view of the entire input. This is achieved with an iterative process using an injection sum with kernel size 1:

$$\begin{aligned} \mathbf{F}'' &= I_1(\mathbf{F}'_{q-1}, \mathbf{F}'_q) + \mathbf{F}_{q-1}, \\ \mathbf{F}'' &= I_1(\mathbf{F}'_{q-i-1}, \mathbf{F}'') + \mathbf{F}_{q-i-1} \quad \text{for } i \in [1, q-1], \end{aligned} \quad (16)$$

where after the last iteration, $\mathbf{F}'' \in \mathbb{R}^{D \times T}$. This differs from the original TDANet, as in the original implementation I_1 is replaced with a different operation. In addition, we have added residual connections from the multi-scale down-sampled features in order to improve training and create a UNet-like [38] structure. Both of these changes greatly improved the consistency and effectiveness of TDANet without increasing the parameters significantly. Note that in Figure 4 we do not show the residual connections as it would make the diagram too complex.

Finally, the feature map \mathbf{F}'' is converted from the hidden dimension D back to the bottleneck dimension B_a or B_v using

a 1D convolution with kernel size 1, and residual connection is added. This result is denoted \mathbf{X}_j for the output of the audio sub-network, and \mathbf{Y}_j for the output of the video sub-network, see Section III-C1.

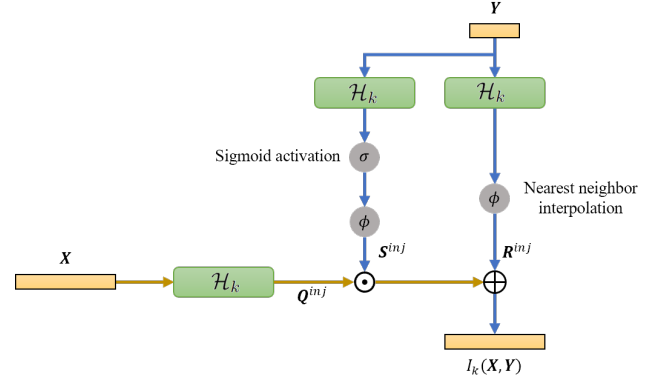


Fig. 6. Architecture for the Injection Sum I_k with kernel size k for two inputs \mathbf{X} and \mathbf{Y} with different temporal dimensions.

V. EXPERIMENTAL PROCEDURES

A. Dataset

Following similar methods [9], [22], [24] in the field, speech separation datasets were constructed from commonly used audio-visual datasets. The speakers in the test set of these datasets did not overlap with those in the training and validation datasets. The raw audio and video frames were obtained using the FFmpeg tool¹. A sampling rate of 16 kHz was chosen, and the model was trained on two seconds of audio and video, equating to audio vectors of length 32,000 and 50 video frames for the 25 FPS video used.

The LRS2 dataset contains a collection of BBC video clips and three separate folders for training, validation and testing purposes. In order to adapt this dataset to an audio-visual dataset, different speakers were randomly selected two speakers at a time. Their audio signals were mixed using signal-to-noise ratios between -5 and 5 dB. The test set is the same as used in previous works [9], [24]. In total, the training set contains 11 hours and the validation set contains 3 hours.

B. Hyperparameter Settings

Following CTCNet [24] we use 16 total layers with 3 fusion layers. The other hyperparameters are defined in Table I. The first block shows the encoder and decoder hyperparameters. The second block shows the audio sub-network hyperparameters. The third block shows the video sub-network hyperparameters.

For training we used a batch size of 4 and AdamW [28] optimization with a weight decay of 1×10^{-1} . The initial learning rate used was 1×10^{-3} , but the learning rate value was halved when the validation data set loss did not decrease for

¹<https://ffmpeg.org/>

TABLE I
TDFNET HYPERPARAMETERS.

Parameter	Value	Description
C_a	512	The audio mixture embedding dimension
K_a	21	The kernel size of the encoder and decoder
S_a	10	The stride of the encoder and decoder
B_a	512	The audio bottleneck out channel dimension
B_v	64	The video bottleneck out channel dimension
D_a	512	The audio sub-network hidden dimension
q_a	5	The audio sub-network up-sampling depth
k_a	5	The audio sub-network kernel size
h_a	512	The GRU hidden dimension
R_a	16	The total number of audio sub-network repeats
D_v	64	The video sub-network hidden dimension
q_v	4	The video sub-network up-sampling depth
k_v	3	The video sub-network kernel size
h_v	8	The number of attention heads
R_f	3	The total number of video sub-network repeats

5 epochs in a row. We also used gradient clipping in order to limit the maximum L_2 norm of the gradient to 5. Training was left running for a maximum of 200 epochs, but early stopping was also applied. Models were all trained on four servers, three containing 8 NVIDIA 3080 GPUs, and one containing 8 NVIDIA 3090 GPUs. The model code is implemented using PyTorch.

C. Loss Function

The loss function used for training is the scale-invariant source-to-noise ratio (SI-SNR) [21] between the estimated and original signals \mathbf{s}_i and $\hat{\mathbf{s}}_i$ respectively for each speaker. SI-SNR is defined as

$$\text{SI-SNR}(\mathbf{s}_i, \hat{\mathbf{s}}_i) = 10 \log_{10} \left(\frac{\|\boldsymbol{\omega}_i \cdot \mathbf{s}_i\|^2}{\|\hat{\mathbf{s}}_i - \boldsymbol{\omega}_i \cdot \mathbf{s}_i\|^2} \right), \quad (17)$$

where $\boldsymbol{\omega}_i$ is the result

$$\boldsymbol{\omega}_i = \frac{\hat{\mathbf{s}}_i^T \mathbf{s}_i}{\mathbf{s}_i^T \mathbf{s}_i}. \quad (18)$$

D. Evaluation Metrics

Following recent literature, the scale-invariant signal-to-noise ratio improvement (SI-SNRi) and signal-to-noise ratio improvement (SDRi) were used to evaluate the quality of the separated speeches. These metrics were calculated based on the scale-invariant signal-to-noise ratio (SI-SNR) [21] and source-to-distortion ratio (SDR) [49]:

$$\begin{aligned} \text{SI-SNRi}(\mathbf{x}, \mathbf{s}_i, \hat{\mathbf{s}}_i) &= \text{SI-SNR}(\mathbf{s}_i, \hat{\mathbf{s}}_i) - \text{SI-SNR}(\mathbf{s}_i, \mathbf{x}), \\ \text{SDRi}(\mathbf{x}, \mathbf{s}_i, \hat{\mathbf{s}}_i) &= \text{SDR}(\mathbf{s}_i, \hat{\mathbf{s}}_i) - \text{SDR}(\mathbf{s}_i, \mathbf{x}), \end{aligned} \quad (19)$$

where

$$\text{SDR}(\mathbf{s}_i, \hat{\mathbf{s}}_i) = 10 \log_{10} \left(\frac{\|\mathbf{s}_i\|^2}{\|\mathbf{s}_i - \hat{\mathbf{s}}_i\|^2} \right). \quad (20)$$

We also consider the number of parameters and the MACs. These metrics are important as they determine the computational complexity and memory requirements of the models. In

both the number of parameters and MACs, a lower value is preferable. For completeness, we also provide PESQ [37] and STOI [45] for the main results tables. For these evaluation metrics, a higher value indicates better performance.

VI. RESULTS

A. Ablation studies

In order to evaluate results faster, experimentation was done using a reduced model setting with 1 fusion layer and 3 audio only layers, totaling 4 layers ($R_f = 1$, $R_a = 4$).

1) *Different Recurrent Operators*: In Table II we examine the effects of using different recurrent operators in the audio sub-network. The transformer, denoted by MHSA, has a good balance between the number of parameters, computational complexity and model performance. It has the lowest number of MAC operations, making the transformer the most efficient choice for audio separation.

A traditional RNN outperforms CTCNet [24] by a significant margin, but lacks the huge performance gains of the transformer. However, we can also see that the RNN model uses significantly less parameters - only 58% of the parameters used by CTCNet. For these experiments, the RNN, GRU and LSTM models all use $h_a = D_a$, the hidden dimension is equal to the input dimension. We found that increasing the hidden dimension to $2\times$ the input dimension barely affected performance, and resulted in a huge increase in the number of parameters.

Moving on to the LSTM [15] and GRU [5] structures, we can see that both offer significant gains over the transformer model, and completely outclass the smaller CTCNet model. As pointed out by Max W. Y. Lam et al. [20], it seems that an RNN based approach is better for audio, which features a high temporal correlation, acoustic signal structure, continuities and sequential nature. Interestingly, even though the GRU model has fewer parameters, it outperforms the LSTM architecture by a significant margin, while also utilizing fewer MAC operations. It appears that for this task, the GRU architecture is the optimal recurrent operator in terms of performance, model size and efficiency. It is worth noting however that the GRU is not without its downsides. The GRU structure does use slightly more memory than the LSTM structure of similar size during training. For readers looking to use larger configurations and who do not care about model size, the LSTM may be the better choice in order to train with a larger batch size.

Table III shows the effect of changing the recurrent operator in the video sub-network, as opposed to the audio sub-network experiments in Table II. Unlike with the audio sub-network, the MHSA clearly outperforms the GRU operator for this medium. These two tables combined show why it is important we use the GRU for the audio sub-network, and MHSA in the video sub-network.

2) *Sharing Parameters*: In Table IV we experiment with sharing the parameters between the audio layers. Both TDFNet and CTCNet achieve a smaller model size compared to VisualVoice [9] by sharing parameters. Specifically, the parameters for the audio sub-network α_j are the same for all j , the

TABLE II
AUDIO SUB-NETWORK RECURRENT OPERATOR ($R_a = 4$, $R_f = 1$).

Model	A_v Module	LRS2-2Mix				Params	MACs
		SI-SNRi	SDRi	PESQ	STOI	(M)	(G)
CTCNet [24]		11.1	11.6	2.90	0.895	6.3	43.6
TDFNet	RNN	11.6	11.9	2.95	0.903	3.6	13.2
TDFNet	MHSA	12.9	13.1	3.08	0.921	4.2	12.8
TDFNet	LSTM	13.4	13.5	3.08	0.928	6.8	15.7
TDFNet	GRU	13.6	13.7	3.10	0.931	5.8	15.0

TABLE III
VIDEO SUB-NETWORK RECURRENT OPERATOR ($R_a = 16$, $R_f = 3$).

β_j Module	LRS2-2Mix				Params	MACs
	SI-SNRi	SDRi	PESQ	STOI	(M)	(G)
GRU	15.3	15.4	3.19	0.942	6.6	47.2
MHSA	15.8	15.9	3.21	0.949	6.5	47.2

TABLE IV
SHARING PARAMETERS IN THE AUDIO SUB-NETWORK ($R_a = 4$, $R_f = 1$)

Audio Shared	LRS2-2Mix				Params	MACs
	SI-SNRi	SDRi	PESQ	STOI	(M)	(G)
X	12.3	12.4	3.01	0.914	22.9	15.7
✓	13.4	13.5	3.08	0.928	6.8	15.7

parameters for the video sub-network β_j are the same for all j and the parameters for the cross-modal fusion sub-network γ_j are also the same for all j . In PyTorch, this structure is realised by instantiating one TDFNet block, and passing the data through this same block R_a or R_f times. In Table IV we use $R_f = 1$, so there are no additional layers for the video sub-network and fusion model to share parameters with. Hence, we experiment with not sharing the parameters between the layers of the audio sub-networks, we are instantiating a new TDFNet block R_a times which results in a huge increase in model size, as seen in Table IV. One might expect this to increase the performance of the model, but we see quite the exact opposite effect. Instantiating new layers results in a large drop in performance. This is likely because parameter sharing allows the TDFNet blocks themselves to act as an RNN-like structure, and so when we stop sharing parameters we lose this important effect.

Table V shows the effect of sharing parameters, this time in the video and fusion layers. If the “shared” column has a “✓” mark, then the parameters for β_j are the same for all j , and the parameters for γ_j are the same for all j : we instantiate one instance of the video and fusion sub-networks, and pass the data through these layers R_f times. We can see that unlike in the audio sub-network, sharing parameters here decreases performance. This is likely due to the nature of the fusion

TABLE V
SHARING PARAMETERS IN THE VIDEO AND CROSS-MODAL FUSION SUB-NETWORKS ($R_a = 16$, $R_f = 3$).

Shared	LRS2-2Mix				Params	MACs
	SI-SNRi	SDRi	PESQ	STOI	(M)	(G)
✓	15.0	15.2	3.16	0.938	4.2	38.6
X	15.3	15.4	3.20	0.943	4.9	38.6

operation. In the first iteration, the video network looks at a pure video feature map generated from the video encoder. In subsequent iterations, there is the combined audio signal and the skip connection. It seems that the model likes to use the subsequent iterations to fuse the different information in different ways, and thus benefits from instantiating separate layers. This comes at the cost of increased parameters, but the increase is small and the performance boost is large. It is also worth noting that the models in Table V use MHSA as the recurrent operator in both the audio and video sub-networks. If model size is of most importance, we can see that we can still achieve over 15dB SDRi and SI-SNRi using only 4.2 million parameters - only 60% of the parameters used by CTCNet.

B. Comparison with the state-of-the-arts

In Table VI we can see how TDFNet compares to the competition. We have chosen the three most interesting results for comparison.

- TDFNet-small is the smaller configuration using only one fusion layer, and three audio only layers. It comes from the last row of Table II. This model is small with outstanding performance and a very low computational cost.
- TDFNet (MHSA + Shared) is the version of TDFNet using MHSA as the recurrent operator in both the audio and video sub-networks, and sharing the video sub-network parameters. It comes from the first row of Table V. This model outperforms the SOTA method CTCNet by a significant margin while using only 60% the number of parameters, and 25% the number of MACs.
- TDFNet-large is the final full-bodied version of TDANet using the GRU in the audio sub-network, and using three separate instances for the video sub-network. It comes from the second row of Table III. This model outperforms

TABLE VI
COMPARISON WITH SOTA METHODS ($R_a = 16$, $R_f = 3$). THE “-” DENOTED RESULTS NOT REPORTED IN THE ORIGINAL PAPERS.

Model	LRS2-2Mix				Params	MACs
	SI-SNRi	SDRi	PESQ	STOI	(M)	(G)
uPIT [19]	3.6	4.8	-	-	92.7	-
SuDORM-RF [47]	9.1	9.5	-	-	2.7	-
A-FRCNN [16]	9.4	10.1	-	-	6.3	-
Conv-TasNet [30]	10.3	10.7	-	-	5.6	-
CaffNet-C [22]	-	10.0	0.94	0.88	-	-
CaffNet-C* [22]	-	12.5	1.15	0.89	-	-
Thanh-Dat [46]	-	11.6	3.1	-	-	-
VisualVoice [9]	11.5	11.8	3.00	-	77.8	-
CTCNet [24]	14.3	14.6	3.08	0.931	7.1	167.2
TDFNet-small	13.6	13.7	3.10	0.931	5.8	15.0
TDFNet (MHSA + Shared)	15.0	15.2	3.16	0.938	4.2	38.6
TDFNet-large	15.8	15.9	3.21	0.949	6.5	47.2

all other models across all metrics using only 47.2 billion MACs for two seconds of audio sampled at 16000 Hz. This represents approximately 30% of the MACs used by CTCNet. At the time of writing this paper, TDFNet is the new SOTA method in audio-visual speech separation. The SI-SNRi score of 15.8 presents a 10% increase in performance compared to CTCNet.

VII. CONCLUSION

Existing multimodal speech separation models are inefficient and have limitations for real-time tasks. We propose a multi-scale and multi-stage framework for audiovisual speech separation based on TDANet and CTCNet. This model can significantly improve the speech separation performance by fusing features of different modalities several times in the fusion stage and influencing the feature extraction network of the corresponding modalities separately. In addition, we explore the impact of different global feature extraction structures on the performance and find that using GRU for sequence modeling can substantially improve the performance and reduce the model computation. Our experiments show that TDFNet outperforms the current state-of-the-art model CTCNet in several audio separation quality metrics while using only 30% the number of MACs.

ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China (grant 2021ZD0200301) and the National Natural Science Foundation of China (grant 62061136001).

REFERENCES

- [1] Triantafyllos Afouras, Joon Son Chung, and Andrew Senior. The conversation: Deep audio-visual speech enhancement. In *Interspeech*, pages 3244–3248, 2018.
- [2] Fahimeh Bahmaninezhad, Shi-Xiong Zhang, Yong Xu, Meng Yu, John HL Hansen, and Dong Yu. A unified framework for speech separation. *arXiv preprint arXiv:1912.07814*, 2019.
- [3] Adelbert W Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128, 2000.
- [4] Céline Cappe, Eric M Rouiller, and Pascal Barone. Multisensory anatomical pathways. *Hearing Research*, 258(1-2):28–36, 2009.
- [5] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, page 1724. Association for Computational Linguistics, 2014.
- [6] Martin Cooke, John R Hershey, and Steven J Rennie. Monaural speech separation and recognition challenge. *Computer Speech & Language*, 24(1):1–15, 2010.
- [7] T Donishi, A Kimura, H Imbe, I Yokoi, and Y Kaneoke. Sub-threshold cross-modal sensory interaction in the thalamus: lemniscal auditory response in the medial geniculate nucleus is modulated by somatosensory stimulation. *Neuroscience*, 174:200–215, 2011.
- [8] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):1–11, 2018.
- [9] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *Computer Vision and Pattern Recognition Conference (CVPR)*, pages 15490–15500. IEEE, 2021.
- [10] Meng Ge, Chenglin Xu, Longbiao Wang, Chng Eng Siong, Jianwu Dang, and Haizhou Li. SpEx+: A complete time domain speaker extraction network. In *Interspeech*, 2020.
- [11] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.
- [12] Reinhold Haeb-Umbach, Shinji Watanabe, Tomohiro Nakatani, Michiel Bacchiani, Bjorn Hoffmeister, Michael L Seltzer, Heiga Zen, and Mehrez Souden. Speech processing for digital home assistants: Combining signal processing with deep-learning techniques. *IEEE Signal Processing Magazine*, 36(6):111–124, 2019.
- [13] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural Computation*, 17(9):1875–1902, 2005.
- [14] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.
- [15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [16] Xiaolin Hu, Kai Li, Weiyi Zhang, Yi Luo, Jean-Marie Lemerrier, and Timo Gerkmann. Speech separation using an asynchronous fully recurrent convolutional neural network. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

- [17] Chanwoo Kim, Anjali Menon, Michiel Bacchiani, and Richard Stern. Sound source separation using phase difference and reliable mask selection. In *International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, pages 5559–5563. IEEE, 2018.
- [18] Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani. Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home. In *Interspeech*, pages 379–383, 08 2017.
- [19] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913, 2017.
- [20] Max WY Lam, Jun Wang, Dan Su, and Dong Yu. Effective low-cost time-domain audio separation using globally attentive locally recurrent networks. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 801–805. IEEE, 2021.
- [21] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr-half-baked or well done? In *International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.
- [22] Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, and Kwanghoon Sohn. Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In *Computer Vision and Pattern Recognition Conference (CVPR)*, pages 1336–1345, 2021.
- [23] Bo Li, Tara N Sainath, Arun Narayanan, Joe Caroselli, Michiel Bacchiani, Ananya Misra, Izhak Shafra, Hasim Sak, Golan Pundak, et al. Acoustic modeling for google home. In *Interspeech*, pages 399–403, 2017.
- [24] Kai Li, Fenghua Xie, Hang Chen, Kexin Yuan, and Xiaolin Hu. An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits. *arXiv preprint arXiv:2212.10744*, 2022.
- [25] Kai Li, Runxuan Yang, and Xiaolin Hu. An efficient encoder-decoder architecture with top-down attention for speech separation. In *International Conference on Learning Representations (ICLR)*, 2023.
- [26] Yuanqing Li, Fangyi Wang, Yongbin Chen, Andrzej Cichocki, and Terrence Sejnowski. The effects of audiovisual inputs on solving the cocktail party problem in the human brain: An fmri study. *Cerebral Cortex*, 28(10):3623–3637, 2018.
- [27] Jiuxin Lin, Xinyu Cai, Heinrich Dinkel, Jun Chen, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Zhiyong Wu, Yujun Wang, and Helen Meng. Av-sepformer: Cross-attention sepformer for audio-visual target speaker extraction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.
- [29] Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation. In *International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, pages 46–50, 2020.
- [30] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, 2019.
- [31] Radek Martinek, Jan Vanus, Jan Nedoma, Michael Fridrich, Jaroslav Frnda, and Aleksandra Kawala-Sterniuk. Voice communication in noisy environments in a smart house using hybrid lms+ ica algorithm. *Sensors*, 20(21):6022, 2020.
- [32] Kana Mizokuchi, Toshihisa Tanaka, Takashi G Sato, and Yoshifumi Shiraki. Alpha band modulation caused by selective attention to music enables eeg classification. *Cognitive Neurodynamics*, pages 1–16, 2023.
- [33] Juan F Montesinos, Venkatesh S Kadandale, and Gloria Haro. A cappella: Audio-visual singing voice separation. In *British Machine Vision Conference (BMVC)*, 2021.
- [34] Ryan J Morrill and Andrea R Hasenstaub. Visual information present in infragranular layers of mouse auditory cortex. *Journal of Neuroscience*, 38(11):2854–2862, 2018.
- [35] Sarah L Pallas, Anna W Roe, and Mriganka Sur. Visual projections induced into the auditory pathway of ferrets. i. novel inputs to primary auditory cortex (ai) from the lp/pulvinar complex and the topography of the mgn-ai projection. *Journal of Comparative Neurology*, 298(1):50–68, 1990.
- [36] Zexu Pan, Ruijie Tao, Chenglin Xu, and Haizhou Li. Muse: Multi-modal target speaker extraction with visual cues. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6678–6682. IEEE, 2021.
- [37] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, volume 2, pages 749–752. IEEE, 2001.
- [38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Interventions (MICCAI)*, pages 234–241. Springer, 2015.
- [39] Mikkel N Schmidt and Rasmus Kongsgaard Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Interspeech*, volume 2, pages 2–5. Citeseer, 2006.
- [40] Khairunisa Sharif and Bastian Tenbergen. Smart home voice assistants: A literature survey of user privacy and security vulnerabilities. *Complex Systems Informatics and Modeling Quarterly*, pages 15–30, 10 2020.
- [41] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11(1-2):13–29, 2005.
- [42] Tao Song, Lin Xu, Ziyi Peng, Letong Wang, Cimin Dai, Mengmeng Xu, Yongcong Shao, Yi Wang, and Shijun Li. Total sleep deprivation impairs visual selective attention and triggers a compensatory effect: Evidence from event-related potentials. *Cognitive Neurodynamics*, 17(3):621–631, 2023.
- [43] Hanne Stenzel, Jon Francombe, and Philip JB Jackson. Limits of perceived audio-visual spatial coherence as defined by reaction time measurements. *Frontiers in Neuroscience*, 13:451, 2019.
- [44] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, pages 21–25. IEEE, 2021.
- [45] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, pages 4214–4217. IEEE, 2010.
- [46] Thanh-Dat Truong, Chi Nhan Duong, Hoang Anh Pham, Bhiksha Raj, Ngan Le, Khoa Luu, et al. The right to talk: An audio-visual transformer approach. In *International Conference on Computer Vision (ICCV)*, pages 1105–1114, 2021.
- [47] Efthymios Tzinis, Zhepei Wang, and Paris Smaragdis. Sudo rm -rf: Efficient networks for universal audio source separation. In *Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020.
- [48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.
- [49] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.
- [50] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A Saurous, Ron J Weiss, Ye Jia, and Ignacio Lopez Moreno. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. In *Interspeech*, 2018.
- [51] Jian Wu, Yong Xu, Shi-Xiong Zhang, Lian-Wu Chen, Meng Yu, Lei Xie, and Dong Yu. Time domain audio visual speech separation. In *Automatic Speech Recognition & Understanding (ASRU)*, pages 667–673. IEEE, 2019.
- [52] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li. Spex: Multi-scale time domain speaker extraction network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1370–1384, 2020.
- [53] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multitalker speech separation. In *International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017.
- [54] Neil Zeghidour and David Grangier. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2840–2849, 2021.