EduBrain 清帆科技

Weekly Tech Salon 论文分享

Google AI

Looking to Listen at the Cocktail Party:

A Speaker-Independent Audio-Visual Model for Speech Separation

「鸡尾酒会效应」：一个音频 - 视觉语音分离模型

Google Research, The Hebrew University of Jerusalem

Content

- 论文项目介绍
- 介绍：鸡尾酒会效应
- 研究成果
- 构建数据集
- 模型细节
- 进一步的考虑 / 模型的借鉴意义 / 相关资源

论文项目介绍

- 实验室 / 机构 : Google Research,
  The Hebrew University of Jerusalem
- 发表时间 : 4 月 11 日 , 2018
- 文章影响力 : ICLR 2018 best paper
- 工作简述 : 谷歌在文中提出了一种基于深度学习网络的用于将单个语音信号与背景噪声、其他人声等混合声音分离开来的音频 - 视觉模型 .
- 关键词 : Audio-Visual, Source Separation, Speech Enhancement, Deep Learning, CNN, BLSTM

介绍：鸡尾酒会效应

- 在嘈杂的环境中，人们非常善于把注意力集中在某个特定的人身上，在心理上「屏蔽」其他所有声音。这种能力被称为「**鸡尾酒会效应**」，是我们人类与生俱来的技能。

  E. Colin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. The Journal of the acoustical society of America 25, 5 (1953), 975–979.

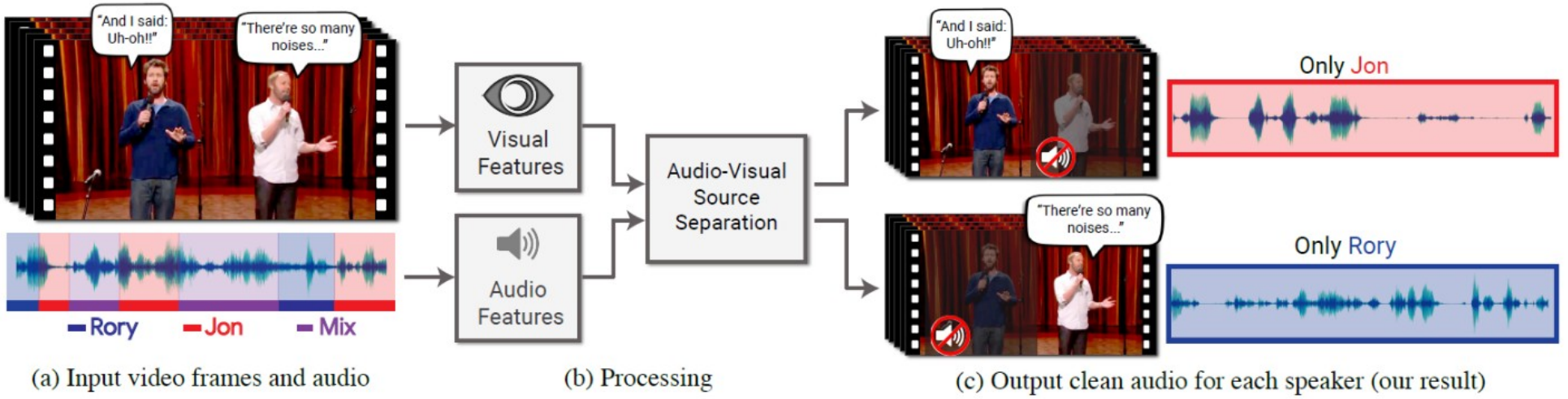
- 研究表明，人可以通过看说话人的脸，有效强化嘈杂环境下对于对方声音的辨识力.

  Elana Zion Golumbic, Gregory B Cogan, Charles E. Schroeder, and David Poeppel. 2013. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". The Journal of neuroscience : the official journal of the Society for Neuroscience 33 4 (2013), 1417–26.

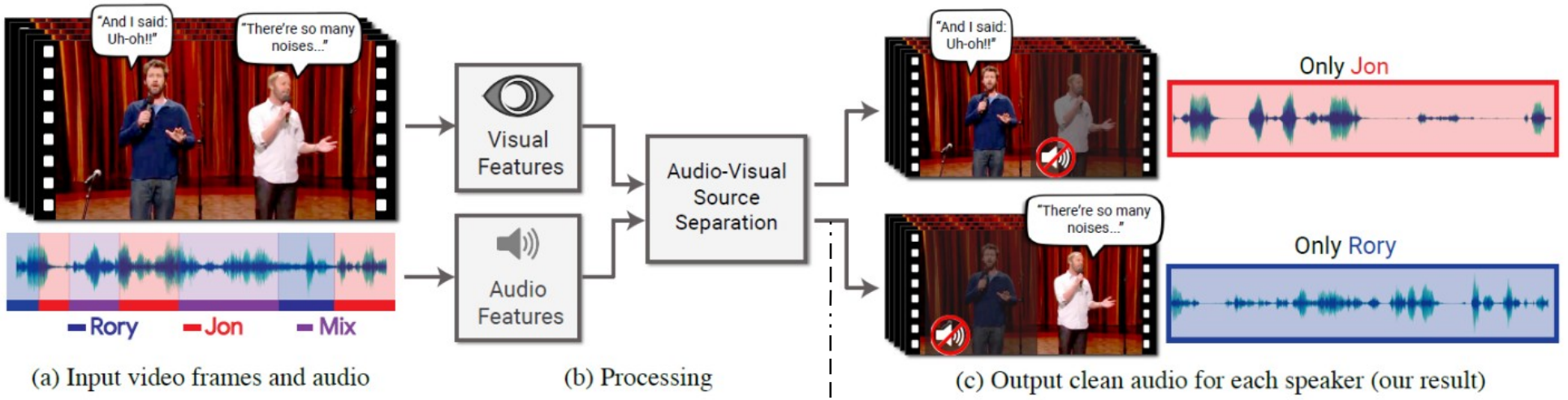
- 本文考虑通过视觉和语音信息来帮助过滤，提取预期的声音.

研究成果

论文展示了一种基于深度网络的模型，该模型整合了视觉信号和听觉信号来解决该任务。视觉特征用于「聚焦」场景中目标说话者的音频(将单个语音信号与背景噪声、其他人声等混合声音分离开来)，以提高语音分离的质量。：



研究成果



(a) Input video frames and audio (b) Processing (c) Output clean audio for each speaker (our result)

Input:

包含多个发言者的视频音频

Processing:

提取视觉特征和声音特征，喂入模型

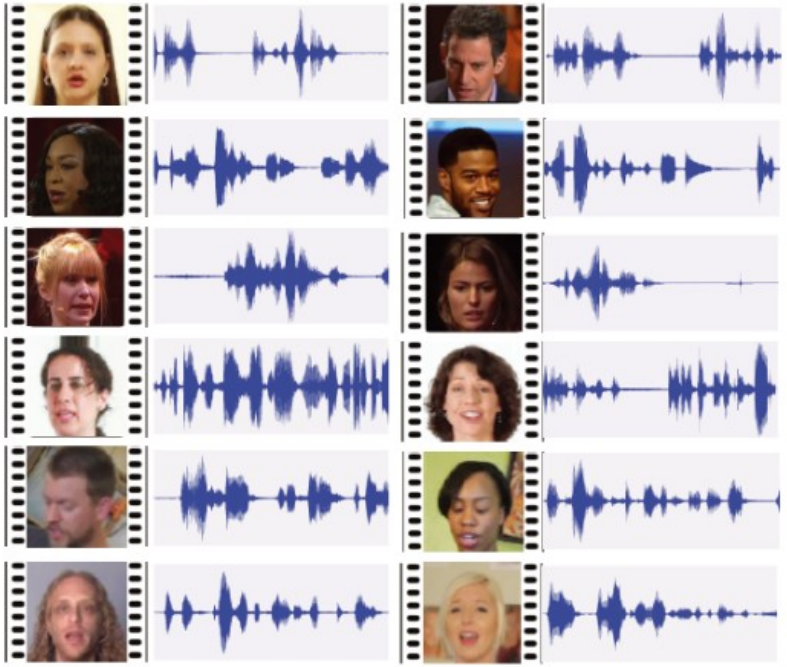
在模型处理中生成噪声掩码

Output:

将输入音频轨道分解成的干净语音轨道，其中每个语音轨道来自视频中检测到的每一个人

AVSpeech dataset 数据集的建立

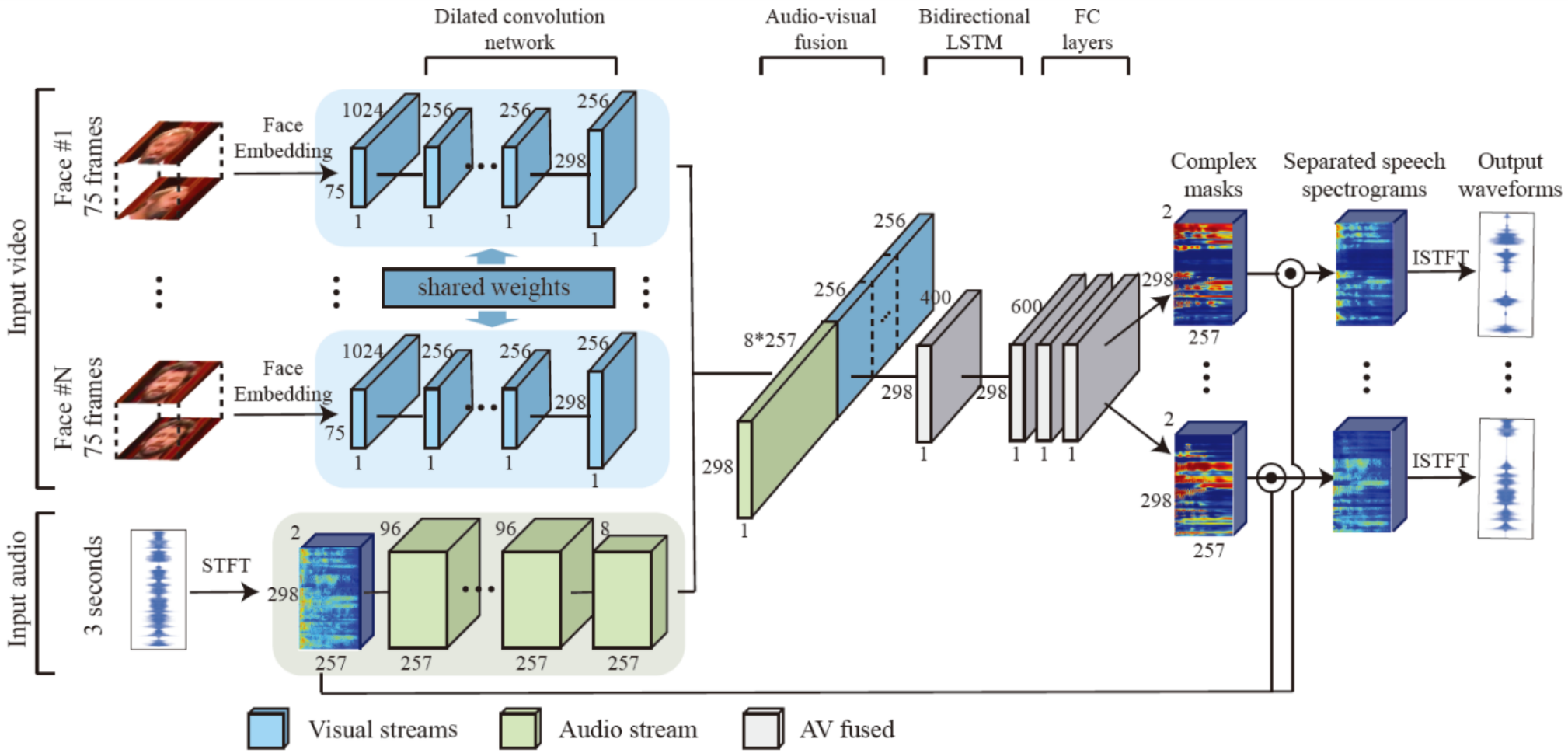
(a) Online videos of talks and lectures we collected



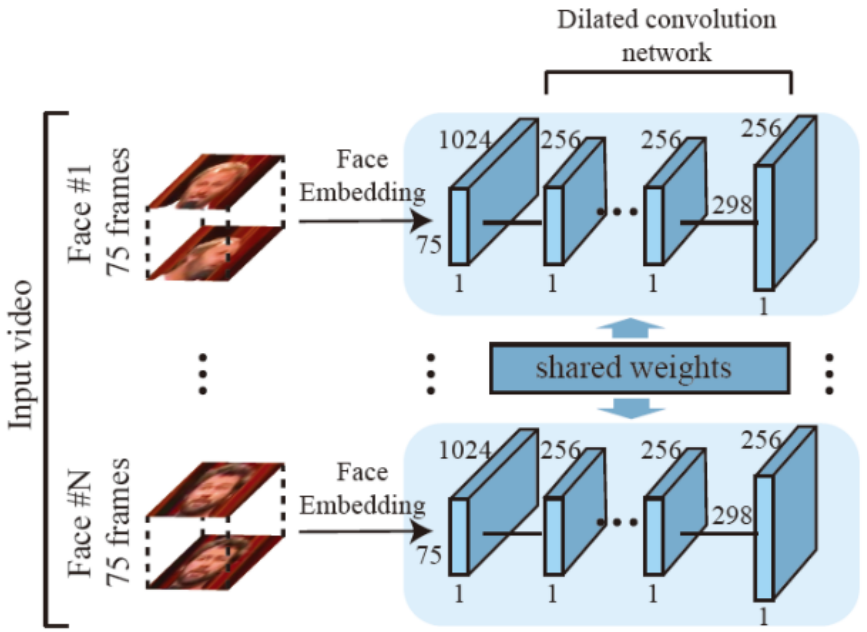
(b) Video segments with localized speakers and clean speech (which comprise our dataset)

- 收集 90,000 个高清的在线演讲或授课视频
- 抽取出其中有清晰语音，并且说话人可见的片段（长度在 3 ~ 10s），得到了 2000 小时的视频片段，每个片段都是清晰无背景干扰的.
- 这些数据囊括不同种族，语言，面部位置等信息.该数据库预备向所有研究机构开放.

Audio-Visual Speech Separation Model 的网络构造



Audio-Visual Speech Separation Model 的网络构造



- off-the-shelf face detector
- use a pretrained face recognition model to extract one face embedding per frame for each of the detected face thumbnails
- Learn a visual feature using a dilated convolutional NN.

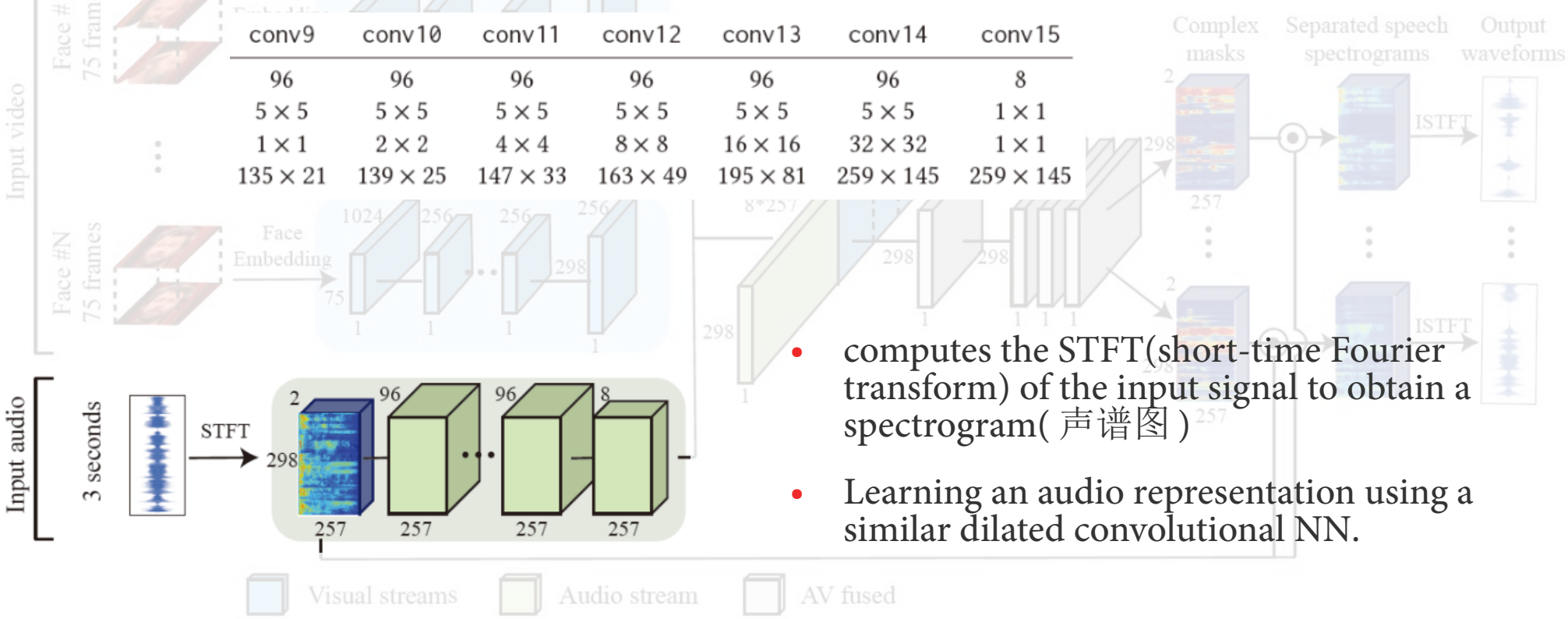
	conv1	conv2	conv3	conv4	conv5	conv6
Num Filters	256	256	256	256	256	256
Filter Size	7 × 1	5 × 1	5 × 1	5 × 1	5 × 1	5 × 1
Dilation	1 × 1	1 × 1	2 × 1	4 × 1	8 × 1	16 × 1
Context	7 × 1	9 × 1	13 × 1	21 × 1	37 × 1	69 × 1

Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. 2016. *Synthesizing normalized faces from facial identity features*. In CVPR'17.

Audio-Visual Speech Separation Model 的网络构造

	conv1	conv2	conv3	conv4	conv5	conv6	conv7	conv8
Num Filters	96	96	96	96	96	96	96	96
Filter Size	1 × 7	7 × 1	5 × 5	5 × 5	5 × 5	5 × 5	5 × 5	5 × 5
Dilation	1 × 1	1 × 1	1 × 1	2 × 1	4 × 1	8 × 1	16 × 1	32 × 1
Context	1 × 7	7 × 7	9 × 9	13 × 11	21 × 13	37 × 15	69 × 17	133 × 19

conv9	conv10	conv11	conv12	conv13	conv14	conv15
96	96	96	96	96	96	8
5 × 5	5 × 5	5 × 5	5 × 5	5 × 5	5 × 5	1 × 1
1 × 1	2 × 2	4 × 4	8 × 8	16 × 16	32 × 32	1 × 1
135 × 21	139 × 25	147 × 33	163 × 49	195 × 81	259 × 145	259 × 145

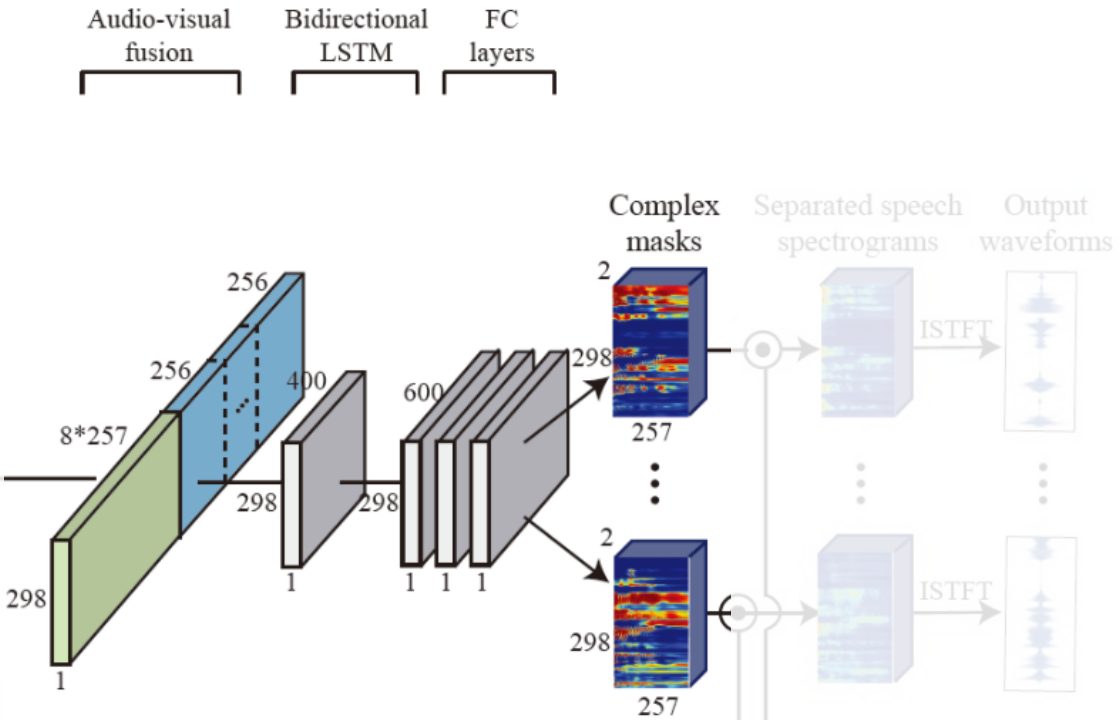


- computes the STFT(short-time Fourier transform) of the input signal to obtain a spectrogram(声谱图)
- Learning an audio representation using a similar dilated convolutional NN.

Audio-Visual Speech Separation Model 的网络构造

- Concatenating the learned visual and audio features.
- Using a bidirectional LSTM and three fully connected layers.
- Output:

  a multiplicative spectrogram mask.

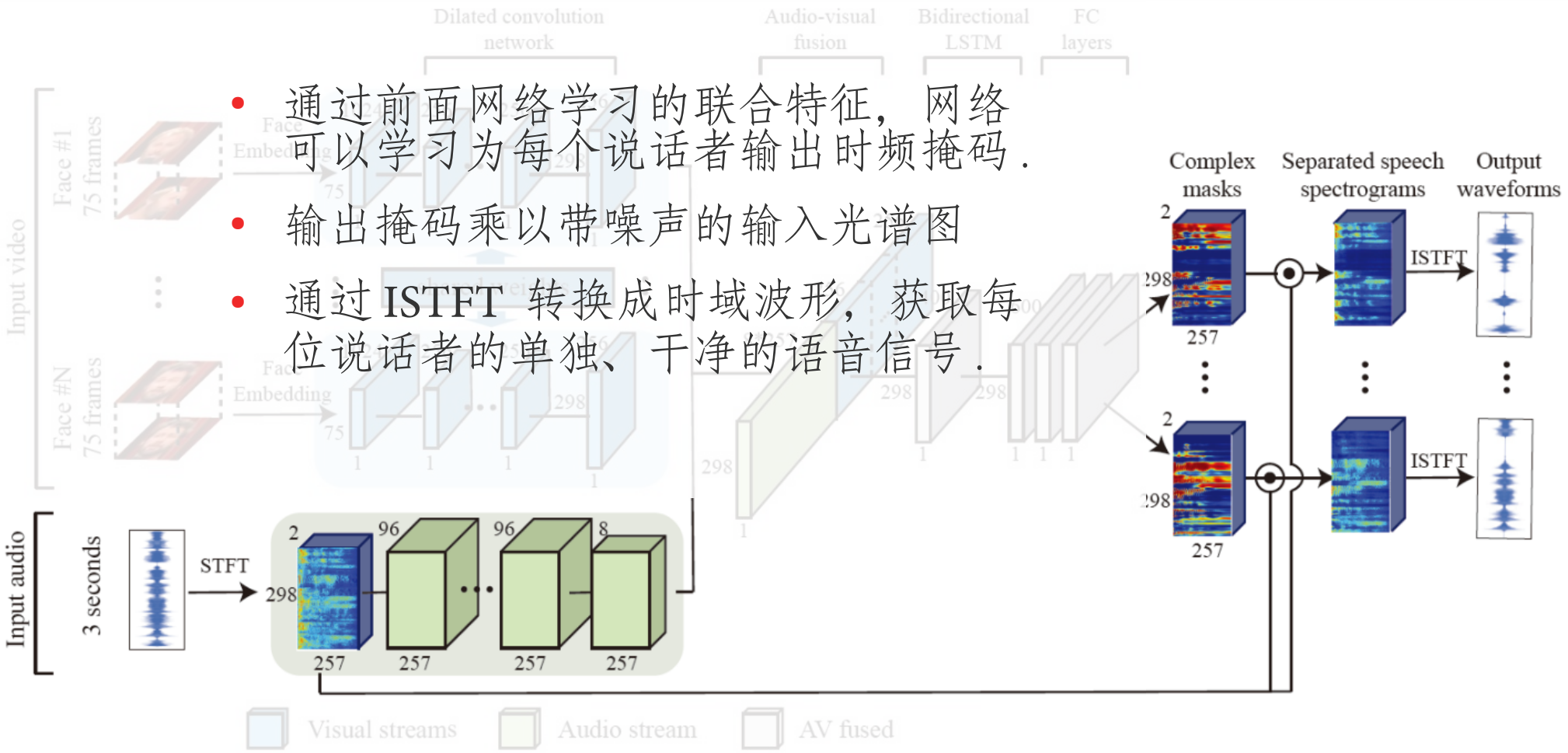


尝试了两种蒙版方法:

ratio mask (RM)
complex ratio mask (cRM)

Audio-Visual Speech Separation Model 的网络构造

- 通过前面网络学习的联合特征，网络可以学习为每个说话者输出时频掩码 .
- 输出掩码乘以带噪声的输入光谱图
- 通过 ISTFT 转换成时域波形，获取每位说话者的单独、干净的语音信号 .



进一步的考虑 / 模型的借鉴意义

- 等待 **Google AVSpeech dataset** 数据集公开
- 考虑利用 **AVSpeech dataset** 数据集，应用在课堂教学视频分析上
- 模型中利用 **STFT** 将声音转化为图像（声谱图），利用卷积神经网络提取特征的方法
- 利用 **Video** 对 **Audio** 进行分离的方法，可以考虑用于说话人识别，通过图像将无监督问题转化为有监督的问题
- 这种方法用途广泛，从视频中的语音增强和识别、视频会议，到改进助听器，尤其适用于有多个说话人的情景

相关资源

- 项目地址：http://looking-to-listen.github.io/
- Google Cloud Vision API
- 一些其他的声音分离的成果

  Deep Clustering

  Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation

  AVDCNN SE model