

硕 士 研 究 生 读 书 报 告



题目 唇形网：基于人物的音频驱动唇部动画

作者姓名 乔 桢

作者学号 21951434

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 2019.12.25

VisemeNet: Audio-Driven Animator-Centric Speech Animation

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Qilei Li

By

Zhen Qiao

Zhejiang University, P.R. China

2019

# 摘要

该篇报告主要对VisemeNet: Audio-Driven Animator-Centric Speech Animation这篇论文进行阐述。该篇论文提出了一个新的基于深度学习的方法，可以直接从输入音频中产生人物的语音运动曲线。所提出的三阶段长短期记忆网络架构主要受心理语言学的观点启发，将语音音频分割成语音组流去构建唇形。其提供了一种自动实时嘴唇同步音频解决方案，可以无缝集成到现有的动画流水线中。对现有的音频数据库中和JALI数据集中的数据进行学习，保证了音素组的完整和发言者语气、口音等多方面的完整性。并通过对真实数据的交叉验证、动画师的评判和编辑、与最近的深度学习嘴唇同步解决方案的视觉比较，以及展示其对于不同说话者和语言的弹性来评估结果。

**关键词：**深度学习，面部动画，神经网络

# Abstract

This report focuses on VisemeNet: Audio-Driven Animator-Centric Speech Animation. This paper proposes a new method based on in-depth learning, which can directly generate voice motion curves of characters from input audio. The proposed three-stage long-term and short-term memory network architecture is mainly inspired by psycholinguistic viewpoints, which divides voice and audio into voice streams to construct horizons. It provides an automatic real-time lip synchronization audio solution, which can be seamlessly integrated into the existing animation pipeline. Learning data from existing audio databases and JALI databases ensures the integrity of phoneme groups, speaker tone, accent and other aspects. The results were evaluated by cross-validation of real data, animator evaluation and editing, visual comparison with recent in-depth lip synchronization solutions, and display of their resilience to different speakers and languages.

**Keywords:** deep learning, human motion, character animation, manifold learning

# 1.简介

计算机生成的人脸的真实性非常重要，娱乐（电影和游戏）、医学（面部治疗和修复）和教育（语言、口语训练和网络助手）等行业都能通过真实地建模和模拟人类面部实现各样的功能和应用。动画角色的不完美模仿会使得观众失去对其的信任和同情，但角色的真实感越强，人们就越不能容忍其有缺陷。毫无疑问的是，动画角色的口语表达是其展现自己表现力的关键。动画电影中通常使用关键帧和动作捕获来刻画生动的人物形象。专业动画师的关键帧非常的生动且可以修改，但耗时耗力。通过动作捕获获取人物的语音行为非常的简单，但很难进行之后的修改和精炼。

该篇论文主要提出了一个基于深度学习的体系结构，以近实时的音频信号直接产生以动画人物为中心的语音动画。通过真实数据的交叉验证来定量评估结果，结合专业动画师的定性评论进一步修改和精化动画输出能力。在不同人、不同性别和不同语言方面的语音训练数据下，该体系结构能够为语音动画提供真正的语言不可知方案。

该篇论文所提出了一个三阶段网络架构，一是学习从音频预测音素组序列；另一个是学习从音频预测重要面部标志的几何位置；最后一个是学习使用音素组和面部标志来产生JALI参数值和稀疏语音运动曲线，从而使面部充满活力。

# 2.具体工作

## 2.1 相关工作

该论文描述了一种实时的语音与动画人物嘴唇匹配的网络框架。当前的视听语音动画的大量研究可以大致分为过程化、行为捕获和数据驱动三类方法。

过程化语音动画将语音分割成一系列音素，通过规则将语音映射到唇形。唇形或可见音位指出在给定音位的顶点处的嘴的形状。基于行为捕获的语音动画将从人类表演者捕获到的运动数据传送到数字人脸上。数据驱动通过使用可变形的、隐马尔可夫和AAM模型从大型语料库中平滑地拼接面部动画数据，以匹配输入语音轨迹。该论文主要使用基于深度学习的语音动画解决方案，通过简单地使用带有文本转录的音频信号，甚至不使用而实现自动唇同步。

该论文介绍的深度学习方法采用三级网络结构：(a)将语音音频分割成与唇形结构相关的语音组流，(b)预测捕捉语音风格线索的面部标志，(c)然后使用提取的音素组、标志和音频特征来生成唇形运动曲线。所有阶段都是实现高性能所必需的。此外，该论文结合不同的视听数据来进行培训。其培训数据包括带有地面真实音素的音频剪辑、带有跟踪标志的视频剪辑，以及带有地面真实景象的3D面部动画。所有阶段都使用多任务学习进行联合训练，以最小化音素组、标记、唇形、共同发音和语音风格参数预测中的错误。

## 2.2 算法设计

该论文设计的算法主要实现图像高质量、动画可编辑、风格感知的实时口语动画。通过输入的音频信号来预测一段时间内的唇形值，以及下巴和嘴唇的参数。

该论文算法包括对音素组、语音风格和唇形的预测。该论文将音频映射到与唇形对应的音素组，发音唇形相同的音素映射到一个几乎相同的唇形。该论文的网络训练成在没有任何文本或语音记录的情况下也能从音频中识别出这些音素组。然而音素组本身没有声音传递的信息，也不能预测视觉，无法表现演讲者的情感，且同一个音素可能在对话中表示发音或尖叫。这些风格属性的音频行为可以通过颌和嘴唇参数得到，与视觉捕获声乐表演的下巴和嘴唇标志密切相关。该论文模型被设计成预测给定输入音频下颚和唇部标志位置随时间的变化。对于唇形预测，结合了音素组、颚部和唇部参数的中间预测以及音频信号本身来产生唇形。

该论文的模型需要来源可靠的多样化训练数据，以适应不同说话人、讲话风格和情感内容等方面。对于大量的语音数据，该论文使用迁移学习训练模型。首先基于从视听数据集中提取的训练音素组和标记，对模型的音素组和标记阶段进行预训练。然后根据预训练好的参数对之后的两个阶段进行初始化，然后对整个模型进行联合训练。在联合训练之前还需要一个音频剪辑的数据集和相关的面部操纵参数流。其中，音素组与唇形高度相关，而标记与颌和唇参数高度相关。因此，在训练唇形阶段，需要使用具有面部操纵参数的示例3D动画的更小数据集来充分推广。此外，该论文采用多任务学习MTL过程来根据音素组、标记、唇形和其他操纵参数预测等多个目标同时训练网络，能够获得较高的性能。

在模型内存方面，该论文采用基于长短期记忆单元LSTM的存储器使神经网络架构能适应模型的所有阶段。该论文认为基于记忆的模型对于从输入音频中正确捕获共同发音和语音上下文非常重要，所以其在输入信号中显式地存储和表示大量上下文。

## 2.3 网络架构

该论文的网络结构采用音频信号作为输入，输出基于JALI的视觉表示。该论文提出的网络架构共有三个阶段：首先通过音素组和标记阶段来处理输入音频，然后通过唇形预测阶段来处理预测的音素组和标记表示以及音频特征。所有的分支都基于将输入信号中的上下文进行编码存储的LSTM单元和将单元的存储器解码为时变预测的全连接层的组合。

该论文对输入的音频信号中的每一帧都提取一个特征向量，用于编码各种功率谱和音频信号特征。提取的特征向量与13个广泛使用的语音识别的Mel频率倒谱系数和26个原始的Mel滤波器组特征相连，这些特征可以很好的识别音频中的情感。另外还有26个频谱子带质心特征，可以很好的识别语音，提高与MFCC结合时的准确性。提取到的65维特征向量会作为输入，传递到网络的三个阶段进行处理。

|  |
| --- |
|  |
| 图2-1 唇形网模型架构图 |

在音素组阶段，会根据当前帧之前的12帧连接的音频特征以及当前帧加上当前帧之后的11帧的音频特征作为输入。由于每隔10ms提取特征，产生的1560维的特征向量，覆盖了240ms窗口中的音频信号信息。这个窗口很好的折中了处理时间和音素组预测经度。特征向量通过三层单向的LSTM单元，分层的更新其内部存储器状态，LSTM单元可以选择将传入特征的表示存储在他们的存储单元中，或者从他们的存储单元中删除表示。通过具有可学习参数的非线性函数，sigmoid函数和双曲正切函数来控制擦除或存储的选择以及输入特征的变换。每一帧中，其最后的LSTM层记忆状态会被解码成音素组的概率预测。总的来说，这个阶段可以被看作一个非线性函数，它考虑到当前帧的音频特征以输出音素组概率。

|  |
| --- |
|  |
| 图2-2 音素组唇形对应表 |

标记阶段将输入的特征向量，通过三个LSTM层并将最上层的存储器解码成38个2D面部标志的稀疏几何，表示每帧的下巴、嘴唇和鼻子配置。标记的目的不在于捕捉特定面部的形态，而是近似地捕捉嘴唇的形状、下颚的位置和平均面部的鼻子。通过预测这些视觉线索可以有效的对讲话风格进行推断。并且，由于语音组和面部下部表面形状是相关的，所以三个LSTM层在音素组和标记阶段之间是共享的。

|  |
| --- |
|  |
| 图2-3 标记阶段结果 |

在唇形阶段，输入参数为音素组、标记位移以及音频特征，得到基于JALI的操纵参数和控制以确定每帧的唇形。由于音素组和标记可能不会完全捕获音频中存在的所有与语音风格相关的信息，所以额外的输入音频特征。

总的来说，该论文主要通过一组三个非线性函数对音素组、标记和唇形进行预测。

## 2.4 模型训练

该论文主要通过迁移学习对模型进行训练。利用公共存储库中的大量带有文本记录的视听数据加上真实音素组，以及跟踪视频中的面部标志作为监控信号，对模型中的音素组和标记阶段进行预训练。之后再用一个小时的音频和JALI操纵参数值的范例流组成的数据集对模型进行微调，来准确的预测唇形。

该论文训练的数据集包括视听数据集和JALI注释的数据集。对于视听数据集，主要来自三个存储库的视听数据，包含了34位发言者以中立的讲话风格所讲的1000个句子的抄本、音频和视频记录，发言者中18位为男性，16位为女性。收录的句子都经过精心挑选，涵盖了英语中常见的音素。该论文对数据集的视频进行了处理，通过DLib和FaceWare Analyzer提取嘴唇、下巴和鼻子的面部标记。这些标记点以中立姿态与虚拟人脸模板进行对齐，并且归一化成相对于人脸位置、方向和大小不变。之后提取每帧相对于平均人脸的训练标志位移。在给定数据集中提供的文本转录后，使用Montreal Forced Aligner将音频与文本对齐，并提取音素以及相关的音素组。

该论文总共两次数据训练。一是预训练，对于给定的N个具有标记位移序列的音频片段以及从训练视听数据集中提取的每帧对应的音素组，估计音素组阶段的解码器参数、标记阶段的解码器参数以及两个阶段共享的LSTM层的参数。对这些参数进行学习，使得预测的音素组尽可能地匹配训练组，预测的标记尽可能地接近训练组，并且预测的标记不会随时间突然变化。二是联合训练，在这个阶段，将音素组和标记阶段的参数初始化为预先训练的值。然后联合估计整个模型的所有参数，包括唇形预测分支的参数，使其基于JALI标注的数据集。

该论文提出的模型总体实现是在Tensorflow上完成，在TitanX GPU上对训练数据集进行测量。

# 3.实验结果

论文中展示了训练和合成角色运动的一些实验结果，并定量和定性地评估了其方法和备选方案。 重点针对BIWI 3D视听数 据集进行研究，来验证论文中的方法和替代方案。

对于定量评估，主要的评价指标为参数的精度和运动曲线的差异。对参数精度的测量基于真值的正确唇形和共同操纵参数的频率。即对于一个给定的二进制变量，指示操纵参数是否在该帧处被激活，并且给定对参数和帧预测的二进制变量。模型预测的精确度即为正确预测操纵参数激活的次数。理想情况下的精确度应该是100%。其次，对于运动曲线差异的测量能有效的评估预测的操纵参数值，如唇形、联合阶段和JALI场的参数，以及这些值在测试框架上平均与真实值之间的绝对差异。其中相应的操纵参数在真实值或预测中是浮动的。由于运动曲线非常稀疏，该论文在评估运动曲线差异时并不考虑非浮动参数，否则零值将主导测量结果。在此阶段，该论文将其模型与多个替代模型进行比较：（1）基于标记：与省略了音素组阶段的模型进行比较，即仅基于标记和音频特征来预测唇形；（2）基于音素：省略了标记阶段；（3）基于音频：消除了标记和音素组两个阶段，仅仅从音频特征直接预测唇形，这意味着没有模型没有经过预训练；（4）没有迁移学习：保留了网络的所有阶段，但数据传送尽在JALI注释的BIWI训练分割上而不是在视听数据集的其余部分上。

|  |
| --- |
|  |
| 图3-1 评估结果 |

对于定性评估，该论文展示了该模型所生成的面部动画结果，以及将其与别的方法所生成的成果进行比较。与该论文的模型相比，先前的方法都未产生可编辑以动画师为中心的唇形曲线或面部动作单元，该论文还演示了涉及不同语言的语音动画的概括。

# 4.结论

该论文提出的基于深度学习的方法，很好的将音频映射到语音运动曲线上。其实现基于手工设计的音频特性，并用模型训练得到的特征进行替换，能够在图像和形状处理管道中执行时改进性能。其次，该模型可以合并一个鉴别器模型，去推断生成的动画的质量，用它的预测提高唇形阶段的性能。最后，本论文的方法只能对面部的下部进行驱动，并未对上部进行学习控制。

该论文的方案实现可能会从根本上改变目前电影和游戏工作室中的面部动画实践，使动画师可以自由地专注于角色表达的创造性和细微的方面。

# 参考文献

[1]Robert Anderson, Björn Stenger, Vincent Wan, and Roberto Cipolla. 2013. Expressive Visual Text-to-Speech Using Active Appearance Models. In Proc. CVPR.

[2]Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. JALI: An Animator- centric Viseme Model for Expressive Lip Synchronization. ACM Trans. Graphics. 35, 4 (2016).

[3]Christoph Bregler, Michele Covell, and Malcolm Slaney. 1997. Video Rewrite: Driving Visual Speech with Audio. In Proc. SIGGRAPH.

[4]Lijuan Wang, Wei Han, and Frank K Soong. 2012. High Quality Lip-Sync Animation for 3D Photo-Realistic Talking Head. In Proc. ICASSP.