CCL24-Eval 任务10系统报告：维沃手语数字人翻译系统

何俊远，刘鑫，杨牧融，李小龙，黄旭铭，滕飞，陈晓昕，付凡
维沃移动通信有限公司
{hejunyuan, liuxin.rgzn, murong.yang, xiaolong.xlli}@vivo.com
{huangxuming, fei.teng, xiaoxin.chen, fan.fu}@vivo.com

摘要

本文介绍了我们在第二十四届中国计算语言学大会手语数字人翻译质量评测中提交的参赛系统。本次评测任务旨在评测手语数字人将汉语翻译成中国手语方面的自然性和准确性。本文介绍的手语数字人翻译系统首先通过手语翻译算法将汉语文本翻译成手语文本，然后将手语文本对应的手语动作单元运用动作融合算法合成为自然、完整的手语数字人动作，同时借助面部驱动算法将口型、表情等非语言元素自然地融入手语合成中，实现带微表情的和唇形同步的手语数字人。最终，我们在官方手语数字人翻译质量的人工评测集上取得了3.513的综合评分，获得了该任务第一名的成绩[1]。

关键词： 手语数字人；手语翻译；动作融合；唇形同步

System Report for CCL24-Eval Task 10: vivo Sign Language Avatar Translation System

Junyuan He, Xin Liu, Murong Yang, Xiaolong Li, Xuming Huang,
Fei Teng, Xiaoxin Chen, Fan Fu
vivo Mobile Communication Co., Ltd
{hejunyuan, liuxin.rgzn, murong.yang, xiaolong.xlli}@vivo.com
{huangxuming, fei.teng, xiaoxin.chen, fan.fu}@vivo.com

Abstract

This paper introduces the competition system we submitted for the sign language avatar translation quality evaluation at the 24th China National Conference on Computational Linguistics. The goal of the evaluation task was to assess the naturalness and accuracy of the sign language avatars in translating Chinese into Chinese Sign Language. The sign language avatar translation system described in this paper first translates Chinese text into sign language text using sign language translation algorithms, then synthesizes the corresponding sign language action units into natural, complete sign language avatar actions using action fusion algorithms, and naturally non-verbal elements such as lip shapes and facial expressions into the sign language synthesis with the help of facial driving algorithms, achieving sign language avatar figures with nuanced facial expressions and synchronized lip shapes. Ultimately, our system achieved a comprehensive score of 3.513 in the official sign language avatar translation quality manual evaluation test set and won first place in this task.

Keywords: Sign Language Avatar , Sign Language Translation , Action Fusion , Lip Sync

[1]https://github.com/ann-yuan/QESLAT-2024

1 引言

手语数字人（Sign Language Avatars，SLA）通过模拟手语动作来实现实时的手语转译，是当前小语种自然语言处理的重要任务之一，它可有效克服听障人士面临的交流障碍，提高该群体的社会参与度和沟通效率。得益于一些手语数据集的开源，比如德国手语数据集RWTH-PHOENIX-Weather 2014(Koller et al., 2015)、美国手语数据集American Sign Language Lexicon Video Dataset(Athitsos et al., 2008)，国内外提出了一些手语生成（Sign Language Production，SLP）模型(Stoll et al., 2018; Saunders et al., 2020; Stoll et al., 2020; Huang et al., 2021)来将手语文本翻译成连续的手语视频流。尽管手语生成技术在全球范围内得到了发展和应用，但特定于中国手语的研究进展却因缺乏专门的数据资源而显得滞后，国内规模较大的手语翻译数据集有CSL-Daily(Zhou et al., 2021)，这种情况也限制了中国手语翻译系统的创新与实现。

当前实现SLA模拟手语动作的核心步骤通常包括汉语转手语、手语合成和通过渲染引擎来获得最终的手语数字人视频序列。由于手语拥有独特的规则和语法体系，这使得汉语与手语之间的相互转译变得极为复杂。此外，手语合成的关键点包括获取手语句子中每个Gloss[2]对应的姿势、关节旋转角度等信息、为相邻的Gloss生成自然的过渡动作和同步展示口型和表情这些面部特征。其中中国手语常用的Gloss所对应的手势、关节旋转角度等信息可以通过动作捕捉技术来获得。虽然这种方法的成本依旧很高，但其优势在于不需要针对特定的手语句子进行数据采集，具有良好的扩展性。此外，采用动作捕捉技术还能确保最终通过渲染引擎合成的SLA效果达到较高的水准。

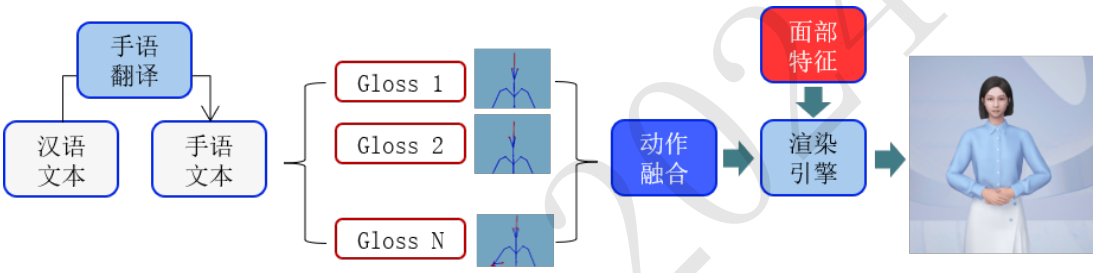

Figure 1: 手语数字人系统流程图

本次评测任务针对手语数字人将汉语翻译成中国手语的语法准确性、自然性、可读性以及文化适应性进行评估。如图1所示，为了保证SLA模拟手语动作的效果，在本次的评测任务中，我们设计了一个包含了手语翻译、动作融合和面部驱动三个主要模块的手语数字人系统。其中手语翻译模块参考了开源的多语言语言模型（Multilingual Random Aligned Substitution Pre-training，mRASP）(Lin et al., 2020)，通过预训练mRASP以及数据增强方法来实现汉语到手语的精准翻译。而动作融合模块借鉴Slot(2007)的工作提出了一种多帧插值平滑算法来使得手语Gloss之间的动作过渡得更加自然流畅，并参考Saunders et al.(2022)的思路设计了一个基于Transformer(Vaswani et al., 2017)的数字人手语合成速度调节模型来改变不同语境下手语Gloss的速度。最后，设计了一个面部驱动模块来把口型和表情等非语言信息融入到手语合成中，提高了手语合成的可懂率。实验结果表明，我们的方法可有效提高SLA模拟中国手语动作的准确性和自然性。

在下文中，我们先后在第二节和第三节介绍参赛系统所涉及到的手语翻译算法和动作融合算法，随后在第四节介绍面部驱动算法。介绍完我们的参赛系统后，我们在第五节和第六节给出实验结果与分析。最后，我们在第七节进行总结并展望未来的研究工作。

2 手语翻译算法

本节主要介绍我们的手语翻译算法。手语翻译是低资源翻译任务，平行语料极度稀缺，标注成本高。我们使用回译(Sun et al., 2020)的方法生成伪双语平行语料，缓解语料稀缺的问题。

[2]类似于汉语词汇，Gloss是手语词汇的唯一标识，如：家-家庭-房子-房，父亲-爸爸，水①，水②。

另一挑战是手语相比汉语更加精炼省略得多，主要是因为手语利用空间特征传达信息。因此，虽然手语的词汇和句子数量较少，但其精炼的细节都在空间特征里，这使得手语和汉语的翻译任务更加复杂。同时由于手语词汇有限，在翻译时针对手语动作需要结合上下文进行转译。基于此我们设计了汉语-手语翻译算法的数据策略和模型策略：

- 使用预训练翻译模型提升低资源翻译任务翻译效果；
- 基于双语平行数据词对齐概率构建中文词与手语词的映射关系；
- 归纳手语语法规则，生成伪平行语料缓解语料稀缺问题；
- 构建模型对手语相同打法候选词进行排序，提升翻译质量。

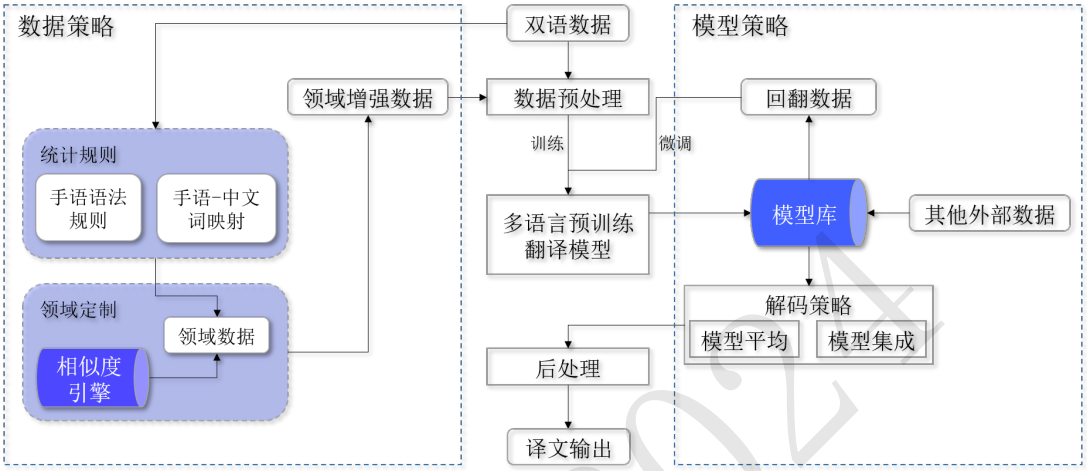

Figure 2: 手语翻译算法方案

本节提到的手语翻译算法使用的是多语言语言模型（mRASP）(Lin et al., 2020)，加入低资源的手语平行语料进行预训练，以便于利用资源较多的语种的先验知识来提升在手语方向上的翻译效果，并基于此预训练模型产生大量伪标签数据作扩充数据，以此训练回翻模型，然后通过迭代的方式持续提升模型性能。图2为我们的手语翻译算法方案。

具体方案的实现上，双语语料是采用了100万语料数据，但数量远远不够支撑模型训练，我们制定了数据策略和模型策略来增强。数据策略指的是从手语和汉语的语法规则以及一些固定的映射关系入手，统计汇总大量规则，基于规则可以用汉语句子构造一些简单的手语Gloss。领域定制指的是如果某些领域的数据特别少，可以定向化召回一些相关领域的示例，构造手语Gloss。模型策略就是利用预训练模型的能力，用100万数据微调之后，通过回翻的手段（汉语转Gloss）来做数据增强。

最终，我们的手语翻译算法在自测集上达到了手语词可懂率：汉语-手语87%，手语-汉语84%；通过句子纠错和上下文语义实现句子词准率70.5%。

3 手语数字人动作融合算法

手语合成的核心是保证语义传达准确，同时手语数字人的动作自然不生硬，贴近真人。这就要求手语句子中每个Gloss的核心动作完整且没有冗余动作，同时相邻Gloss对应的动作衔接自然。此外，不同语境下的同一个或不同的手语动作都要有节奏感，贴近真实的用户场景。为了实现这些目标，必须开发合适的手语合成算法来确保手语动作细节处的平滑过渡和手语动作节奏的恰当调整。本节将介绍手语数字人系统所涉及的动作融合算法，包括多帧插值平滑算法和基于Transformer(Vaswani et al., 2017)的数字人手语合成速度调节算法。

3.1 多帧插值平滑算法

动作混合（Motion Blending）通常在动画、游戏开发和电影制作中使用，指的是将两个或多个动画片段平滑地过渡合并，以创造出一个连贯且自然的动作序列。Slot(2007)提出了一种通

用的方法来混合两种动作使这两种动作之间的变化尽可能逼真，即利用时间扭曲自动确定两种动作之间的时序，并应用动作对齐来控制动作的方向。我们借鉴该方法提出了一种多帧插值平滑算法来为手语句子中相邻的Gloss生成连贯且自然的过渡动作。

构建手语动作库：针对每一个中国手语常用的Gloss，先通过动作捕捉技术从专业的手语老师中收集到该Gloss每一帧的姿势、关节旋转角度等信息。接着人工删除每个Gloss的冗余帧，即完整保留每个Gloss帧序列的核心动作和适当保留Gloss帧序列前后的部分过渡动作。最后这些处理好的Gloss帧序列构成一个完整的手语动作库。

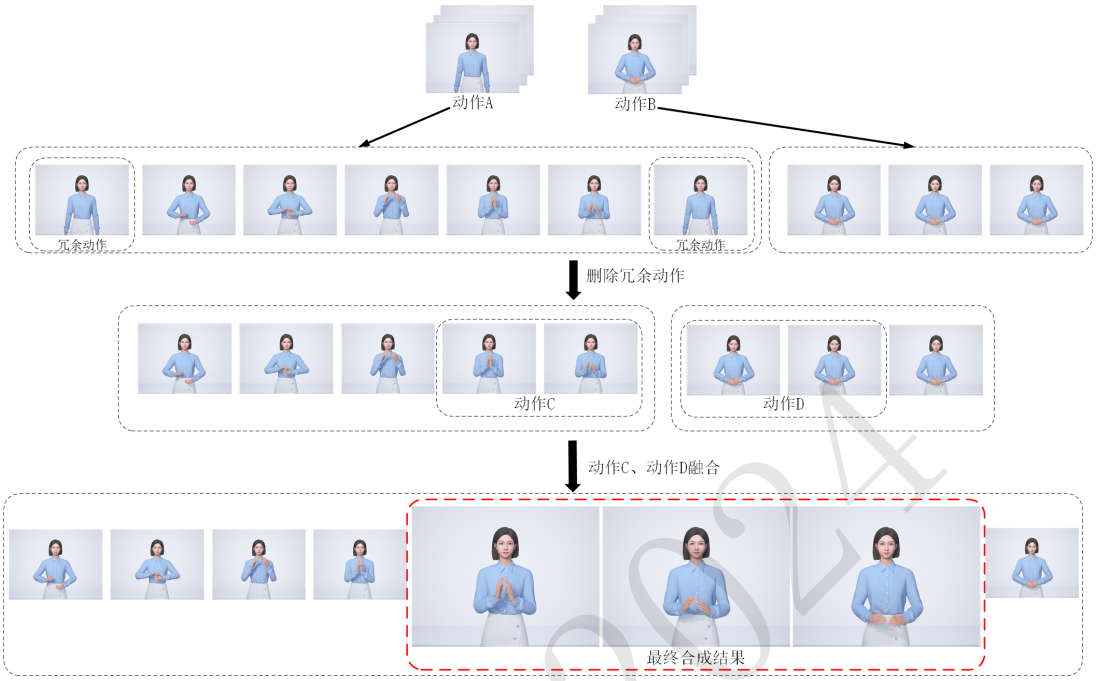

Figure 3: 多帧插值平滑算法流程图

	C+0	C+1	C+2	C+3	C+4	C+5
D+0	20069.	21106. 5	21902. 6	22374. 4	22476. 6	22409. 5
D+1	21148. 4	20894. 9	21487. 6	21860. 8	21949	21969
D+2	19907. 8	20409.	20838	21126. 4	21216. 7	21299. 5
D+3	19699. 3	20018. 5	21227. 1	21573.	21649. 6	22074. 2
D+4	19779. 5	19952. 3	20164. 5	20367. 8	20440. 2	20597. 8
D+5	19714. 6	19799. 5	19957. 2	20139. 3	20212.	20386. 2
D+6	19459. 5	19511. 9	19654. 7	19835. 9	19912. 9	20088. 1
D+7	19282	19326. 8	19470. 2	19656. 3	19737. 4	19908. 2

Figure 4: 动作C第m帧与动作D第n帧所有骨骼关节的空间距离累加结果

确定动作混合轨迹：在第2节中，通过手语翻译模块，我们将汉语文本转换为相应的手语文本Gloss。对于每个得到的手语Gloss，我们从预先构建好的手语动作库中选取相应的帧序列来合成动作。如图3所示，当合成手语动作时，我们需要组合两个帧序列：前者为帧序列A，后者为帧序列B。为了确保合成动作准确传达手语语义，我们提取帧序列A末尾的一定长度的帧作为序列C，并从帧序列B的开头选取一定长度的帧作为序列D。接下来，我们利用BVH（Biovision Hierarchy，BVH）动作文件中的数据，计算帧序列C和帧序列D中每一帧的骨骼关节空间坐标。这些数据包括骨骼关节在空间中的初始姿态坐标和每一帧对应的旋转信息。通过计算骨骼关节的空间位置，我们可以进一步计算在两个动作帧之间的骨骼关节间的空间距离。该空间距离通过累加所有关节之间的距离来计算得出。最后可获得如图4所示的表格，表内单元格数据为动作C第m帧与动作D第n帧所有骨骼关节的空间距离累加的结果。为确保合成手语动作的流畅性和连贯性，必须对参与合成的动作帧进行细致选择。具体而言，在融合动作A和动作B时，每一帧融合生成的动作帧都需要从动作C和动作D的帧集合中挑选出一对空间

位置差异最小的对应帧。而融合生成的每一动作帧，主要将从动作C和动作D中选取出来的动作帧进行四元数差值，差值不是直接计算两帧的平均值，而是需要通过弹簧阻尼的方式去实现，类似于一个缓入缓出的曲线，最终合成的动作中，动作C的权重逐渐由1到0，动作D的权重逐渐由0到1。该操作旨在找到一条从动作C初始帧$(C+0)$至动作D终末帧$(D+n)$的融合路径，沿途保持空间变化的最小化。为平衡合成动作对原始动作A和B的依赖程度，并避免偏向任一动作，我们对图4所示的帧选择策略施加一个约束：即限制连续在一直线方向（横向或竖向）上选择的单元格数量，如最多允许连续选择3个单元格。该规定需灵活应用，以满足实际动作合成的需求，并最终实现在终结帧$(D+n)$上获得平滑的合成轨迹。

3.2 基于Transformer的数字人手语合成速度调节算法

对于上述依靠每个手语Gloss对应的手语动捕数据来进行数字人手语合成的策略，优点是只需获取常见的手语Gloss的动捕数据，此时每个手语Gloss对应的动作的速度是相对一致的。而对于不同手语句子中的同一个手语Gloss对应的动作，若其速度可根据手语句子的特定语境进行动态调整来使得数字人的动作更加自然和容易理解，则可以缓解经过多帧插值平滑算法合成的手语数字人视频序列整体韵律节奏单一的问题。

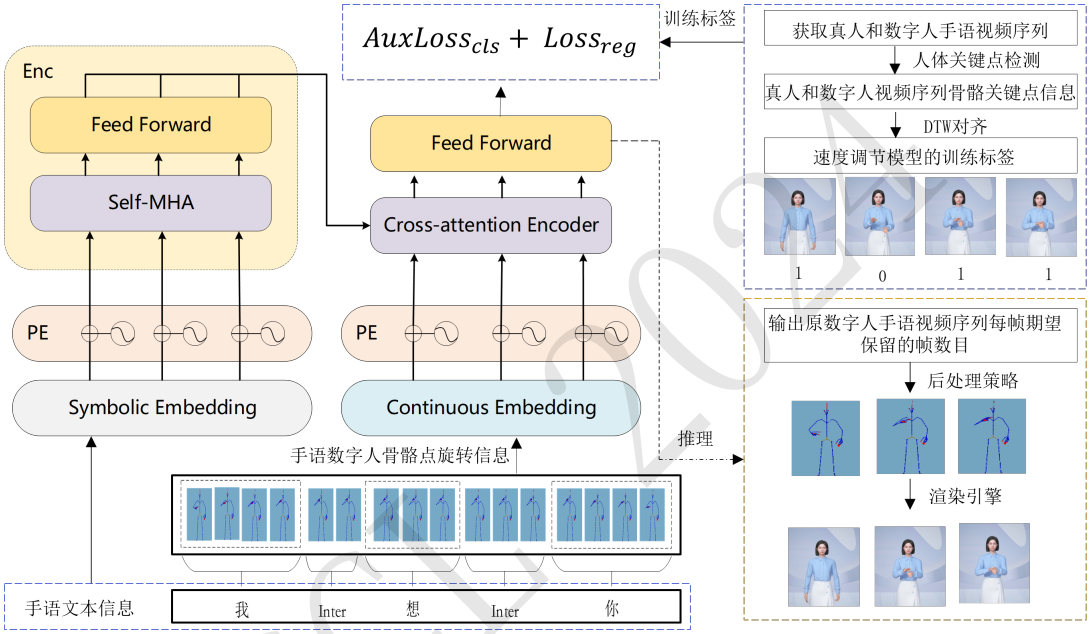

Figure 5: 数字人手语合成速度调节模型架构图

我们参考Saunders et al.(2022)的思路提出了一种基于Transformer的数字人手语合成速度调节算法，使得合成的数字人视频序列拥有贴近真实场景的韵律节奏，以促进用户的理解。其模型架构如图5所示，训练模型时的输入包含手语文本中的Gloss信息、手语数字人视频序列中的骨骼点旋转信息和手语数字人视频序列的训练标签。速度调节模型会先通过Cross-attention Encoder模块来融合手语文本信息和手语数字人视频序列中每一帧所包含的骨骼点旋转信息，它能辅助模型学习到手语句子中关键的上下文信息。最后在预测头部分，包含了一个回归损失和一个分类损失。

获取训练标签： 具体的，如图5所示，假设通过多帧插值平滑算法获得手语文本Gloss对应的手语数字人视频序列为X_u，而该手语句子对应的真人视频序列为T_r，首先使用开源的BlazePose(Bazarevsky et al., 2020)人体骨骼关键点检测技术，从X_u和T_r的每帧中提取出手部和头部的三维骨骼关键点信息。接着进一步利用这些关键点信息和每个Gloss在X_u及T_r中的起始和结束时间，通过动态时间规整（Dynamic Time Warping, DTW）(Müller, 2007)算法对齐两个视频序列。通过这种对齐，我们可以精确匹配X_u中每一帧在真实韵律节奏下所对应的帧数目K_t。K_t的值是一个离散标量，其中0表示该帧被忽略，1表示保留该帧，而大于1的值表示此时需要K_t帧来调整X_u的韵律。通过上述处理，获得的K_t值序列作为速度调节模型训练的标签。

设计类别感知均方误差损失函数: 我们提出了一个新的损失函数，称为类别感知均方误差损失函数（Classification Aware MSE Loss），用于动态调整多元回归任务中不同特征的权重，可有效缓解特征不平衡的问题。假定W_j表示权重占比，j表示X_u中的第j帧，C_j代表分类头输出的第j帧的分类预测值，即预测保留第j帧的数量，T_j是第j帧的真实标签，即实际需要保留的第j帧数量，则有

$$W_j = \frac{\max(C_j, T_j) + 1}{\min(C_j, T_j) + 1} \tag{1}$$

在模型训练过程中，由于C_j的输出结果是变化的，当预测值C_j接近真实标签T_j时，权重W_j接近1。若预测值C_j与真实标签T_j之间的差距增大，W_j则相应增大，以便使模型在训练过程中更加关注那些与真实标签差异较大的特征。因此，W_j能够在训练中动态调整不同特征的权重占比。

对于分类子网络，为了获取更准确的分类预测值C_j，我们采用了标签平滑的交叉熵损失函数$AuxLoss_{cls}$，其中K是类别的总数，n表示X_u中包含的总帧数，t_{ij}是第j帧真实标签对应的独热编码中的第i个元素，c_{ij}是模型预测得到的概率分布中第j帧的第i个元素，ϵ是标签平滑的平滑参数，则有

$$AuxLoss_{cls} = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{K}\left((1-\epsilon)t_{ij} + \frac{\epsilon}{K}\right)\log(c_{ij}) \tag{2}$$

对于回归子网络，我们采用了类别感知均方误差损失函数$Loss_{reg}$，其中n表示X_u中包含的总帧数，R_j表示回归头输出的第j帧的回归结果，即预测需要保留的第j帧数量，则有

$$Loss_{reg} = \frac{1}{n}\sum_{j=1}^{n} W_j \cdot (R_j - T_j)^2 \tag{3}$$

模型推理时的后处理: 在推理阶段，模型将输出原手语数字人视频序列每一帧需要保留的帧数目K_p。为了确保动作过渡的平滑度，采用了如下的后处理操作：

- 当$K_p = 0$时，相关帧将被舍弃；
- 当$K_p = 1$时，该帧被直接保留；
- 当$K_p = 2$时，保留当前帧，并通过对当前帧与后续帧进行非线性插值操作以生成一个新的帧；
- 当$K_p = 3$时，首先对当前帧与其前一帧进行非线性插值以产生一个新的帧，继而保留当前帧，并对当前帧与后续帧执行非线性插值操作以获取另一新的帧；
- 当$K_p \geq 4$时，首先对当前帧与其前一帧执行非线性插值以产生一个新的帧，随后复制当前帧两次，并对当前帧与后续帧执行非线性插值操作产生一个额外新的帧。

最后，我们使用渲染引擎对经过后处理的手语数字人视频序列中每一帧对应的动捕数据进行渲染处理，生成了流畅且节奏感恰当的手语数字人视频序列。

4 手语数字人面部驱动算法

本节主要介绍我们的面部驱动算法。手语表达要实现高准确率和高可懂率，除了保证手语动作的正确和流畅，口型和表情等非手控信息也尤为重要(Von Agris et al., 2008)。打手语时附有口型，口型准确，能辅助听障人士理解动作；打手语时口不动，让用户感受不到数字人在与他聊天。打手语时准确呈现表情，表情生动，适当夸张，提升真实感；打手语时无表情，形象呆板，容易让听障人士产生歧义。

本节介绍我们所设计的面部驱动算法，图6为具体算法架构。涉及非语言信息，主要包括口型驱动和表情驱动两个模块。在高质量口型合成数据缺乏的情况下，我们舍弃了Audio2Face(Karras et al., 2017; Tian et al., 2019)的端到端方案，采用参考Jali(Edwards et al., 2016)的可解释性强，可配置化程度更高Pipline方案。

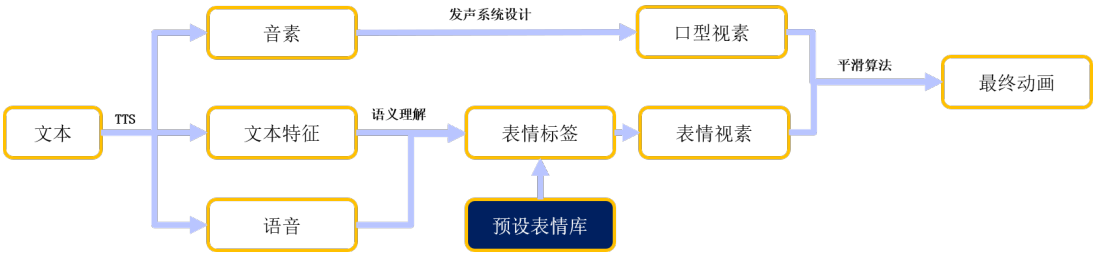

Figure 6: 面部驱动算法架构

4.1 口型驱动算法

我们将文本作为口型驱动的源头。将文本时间序列转换成音素时间序列，并通过口型驱动算法，生成嘴型和面部微表情系数，再通过驱动引擎实现数字人的面部动作的精准驱动。该此方法与语音无关，只与文本内容相关，不受语音特性变换影响。为了实现手语文本语音和数字人口型播报的同步，需要获取时间同步的口型BS（BlendShape，BS）参数以驱动数字人。

将第2节自然文本经过手语翻译算法得到的手语文本Gloss，再经过vivo自研文本语音合成TTS系统生成语音和对应音素，保证语音和音素序列是时间同步的。文字中每个字的多个音素p，以及每个音素的持续时间共同构成语音文字对齐信息。

对音素序列进行分组，同时每组音素设定一个重要性权重w_1，让对口型影响大的音素具有更大的权重。将语音文字信息中的每个音素持续时间进行分析，将输入文本时长按时间间隔T等分为多个。

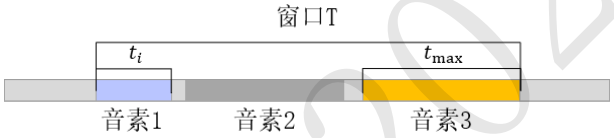

Figure 7: 音素窗口示意图

图7为音素窗口示意图。在时间窗口T中计算音素的密集程度。第i个音素的密集程度w_2计算方式为

$$w_{2i} = (\frac{t_i}{t_{max}} + \frac{t_i}{T})/(n+1) \tag{4}$$

其中t_i表示当前第i个音素持续时间，t_{max}表示时间窗口T中最长的音素持续时间，n表示时间窗口T中。通过这个权重，可以丢弃密集度高但是对口型影响较小的音素，避免唇形的抖动问题。音素参考拼音类型可以分为声母和韵母，其中韵母可以分为单韵母、复韵母、前鼻韵母、后鼻韵母、整体认读音节、三拼音节。对三拼音节和整体认读音节进行拆分，拆分为前四种韵母的组合。最后音素序列分为声母、单韵母、复韵母、前鼻韵母、后鼻韵母，同时每组音素设定一个重要性权重，让对口型影响大的音素具有更大的权重。一般的，对于重要性权重，设定声母、单韵母、复韵母、前鼻韵母、后鼻韵母的权重分别为（1.0，0.9，0.6，0.5，0.5）。最终根据语音对齐信息得到重要性权重w_1和音素的密集度权重w_2。

然后进行数字人口型映射，将音素和数字人口型BS参数进行逐一映射，设计合理的数字人发音系统，使得数字人单一口型和现实人物一致。为了方便起见，我们将一个口型对应的一组BS参数叫做视素v(Bear and Harvey, 2017)，则可以将音素序列转换成视素序列。由于音素的个数是有限且独立的，我们首先使用常见的面部工具（如LiveLinkFace），通过人工采集获取不精确的音素到BS参数映射关系。然后专业的3D建模工程师人工调整虚拟人的BS参数，完成音素到视素的精确表达。最终将对应的音素序列P转变为驱动参数序列V，对于每帧i来说，视素驱动参数：

$$v_i = min(S(p_i) * w_{1i} * w_{2i}, 1.0) \tag{5}$$

其中w_{1i}为音素的重要性权重，w_{2i}为音素密集度权重，S为音素到视素的映射关系。根据前述得到的离散音素序列，得到相应的视素参数序列。

为了解决离散视素间的平滑过渡，我们使用SG（Savitzky-Golay，SG）时间序列平滑算法(Schafer, 2011)对视素的不同参数分别进行插值和平滑，SG平滑算法是在一个滑动窗口内，进行多项式拟合。该方法可以实现自然的口型切换，更良好的交互体验。

字平滑策略：将每个汉字持续的时间（即一组音素对对应的时序序列）作为一个平滑窗口，将V分为多个不同的词窗口V_i，对于每个V_i应用SG算法，保证每个字对应的口型自然。

句平滑策略：将每个字平滑后的视素序列再次应用SG算法对整句话进行平滑，保证唇形运动自然。

$$V = SG\underbrace{((SG(v_1), \ldots, SG(v_i))}_{m} \tag{6}$$

其中m表示该序列中字的个数。最终得到平滑的视素序列V_s，用于驱动手语数字人口型。根据文本信息，生成数字人模型的口型变化序列和语言，然后驱动数字人模型的口型变化，使其与现实人物的语音保持一致。

4.2 表情驱动算法

对文本特征进行提取后，通过语义理解得到情感标签。我们提前制作了数字人通用的预设表情库，将得到的情感标签与预设表情进行一一映射。表情驱动算法同样是基于视素进行数字人面部微表情的驱动，预设表情同样是可以认为是视素。将表情视素和口型视素直接相加，再通过平滑算法进行平滑，即可实现带微表情的唇形同步数字人面部驱动。

最终我们的面部驱动算法集成了口型驱动和表情驱动能力，支持中英文所有发音的口型，支持13种表情，主观评测口型一致性大于90%。

5 自测结果

本节主要介绍我们的自测方法和评测结果。手语是小语种，语料库和相关公开评测集有限，难以自动化评测。我们采用自构建手语语料库的形式，并聘请相关手语专家进行人工评测，确保我们的手语数字人能够较好的满足听障人群的需求。

5.1 评测方式

为了对整个手语数字人将汉语翻译成中国手语方面的自然性和准确性进行评估，我们先采集了约6000句通用场景下包含书面语文本及其对应的手语转写文本的句子，然后精选其中的约1500句包含常用Gloss的句子，覆盖日常生活、工作学习等场景，构成自测专用的手语语料库，并基于该语料库对我们系统的手语合成结果进行人工评测打分。我们邀请多位手语专家进行人工评测，以手语语法的准确性、表达的自然性和可读性以及是否满足听障人士理解为主要标准。

通过我们的手语数字人系统获取到手语语料库中每个手语句子对应的手语数字人视频，统计每一个手语数字人视频翻译正确的词占该手语句子全部评测词的比值，取平均后作为手语词可懂率；统计每一个手语数字人视频的翻译得分，取平均后作为手语句可懂得分。

5.2 评测标准

对于汉语-手语翻译的评测标准，参考中英翻译的评价指标，我们定义0-5分的可懂率，我们认为3分是一个大致能够还原原意的标准。以下是具体的评测细则：

- 0分：合成的手语与原文完全不对应；
- 1分：看了手语不知所云，仅个别词语合成正确，无语义和逻辑；
- 2分：手语与原文小部分符合；有漏译、误泽或严重语法错误；
- 3分：手语大致表达了原文的意思，但对译文整体理解影响不大；
- 4分：手语基本较流畅地表达了原文的意思，但不影响整体理解；
- 5分：手语准确且完整地表达了原文信息，表达流畅自然。

第二十三届中国计算语言学大会论文集，第382页-第392页，太原，中国，2024年7月25日至28日。
卷3：评测报告

5.3 评测结果

我们基于自构建的通用场景评测集，并邀请多位手语专家进行人工评测，最终的评测结果准确率指标为：

- 手语词可懂率96.82%，手语句可懂得分4.13分（满分5分）；
- 自然程度主观评价：流畅度与真实感优秀；
- 口型与表情：支持中英文所有发音的口型，支持13种表情。

6 官方结果与分析

CCL24-Eval 手语数字人翻译质量评测以手语语法的准确性、表达的自然性和可读性以及是否满足聋人理解为主要标准，综合考虑手势清晰度、流畅性、与汉语原文的语义一致性等[3]。手语数字人翻译质量的人工评测包括四个主要指标：手语语法准确性、自然性、可读性以及文化适应性。综合得分是根据每个单项指标的得分与其相应的权重系数计算得出的加权和。评测重点关注手语数字人在准确表达手语的能力，强调自然性和可读性，同时考量其文化适应性。

Table 1: CCL24-Eval 手语数字人翻译质量的综合评测结果

排名	队伍	得分	单位
1	维沃手语数字人团队	3.513	维沃移动通信有限公司
2	team 1	2.447	-
3	team 2	2.119	-
4	team 3	1.806	-

Table 2: CCL24-Eval 手语数字人翻译质量不同维度的评测结果

专家组评分				
队伍	准确性	自然性	可读性	文化适应性
维沃手语数字人团队	3.50	3.75	3.43	2.36
team 1	2.21	2.36	2.00	1.79
team 2	2.61	2.25	2.21	2.07
team 3	1.04	2.14	1.68	1.75
采集组评分				
队伍	准确性	自然性	可读性	文化适应性
维沃手语数字人团队	3.39	3.43	2.86	2.43
team 1	1.11	2.07	1.86	1.29
team 2	2.14	2.07	1.96	1.11
team 3	0.25	1.79	1.54	0.82
普通组评分				
队伍	准确性	自然性	可读性	文化适应性
维沃手语数字人团队	4.14	4.14	3.75	3.11
team 1	1.75	2.89	2.43	2.50
team 2	2.93	2.89	2.89	2.57
team 3	1.79	2.86	2.54	2.04

[3]https://github.com/ann-yuan/QESLAT-2024

表1展示了CCL24-Eval 不同队伍（前4名队伍）的手语数字人翻译质量综合评测结果，其中，我们所开发的手语数字人系统在本次评测中表现卓越，其成绩较第二名领先超过1分，说明本文方法在手语合成任务上有着更为优越的性能，能保证手语数字人将汉语翻译成自然、准确和可读性强的中国手语，并且能被更多的聋人群体所理解和接受。

表2为CCL24-Eval 官方提供的专家、采集和普通组在手语语法准确性、自然性、可读性以及文化适应性四个评测指标上不同队伍的手语数字人翻译质量评测结果，我们提出的手语数字人系统在专家、采集和普通组所评测的准确性、自然性、可读性以及文化适应性这四个指标均取得了本次评测的最佳成绩。其中，本文提出的手语翻译算法可获取更符合中国手语词序规则的手语文本，而动作融合算法则将手语文本对应的每个手语动作单元融合成流畅和手势形态更精准清晰的手语数字人动作，同时面部驱动算法将非言语元素自然地融入翻译中，这些都使得我们的手语数字人系统在手语语法准确性、自然性和可读性均取得较好的评测成绩。但在文化适应性上相比其它指标有1分左右的差值，具体原因是本文提出的手语翻译算法在翻译过程中有时并未考虑到文化差异和特定的社会语境，如礼貌用语、行业术语等，同时我们提出的表情驱动算法当前只支持13种表情，有时无法映射出某些手语句子对应的情感色彩。而较低的文化适应性也会影响到整个手语数字人系统在自然性和可读性上的表现，比如部分面部表情未自然地融入翻译中时会显现出生硬、不自然的状态和在不同语境下的适应性未被考虑时会影响其可读性。

7 结语

本文以提高听障人士和健听人士的双向沟通效率为动机，为广大听障群体提供沟通上的便利，提出了结合手语翻译算法、动作融合算法和面部驱动算法的手语数字人翻译系统。本文提出的系统在手语专家的评测下达到了手语词可懂率96.82%，手语句可懂得分4.13分（满分5分）的优秀性能，同时手语语义传达较为准确，手语数字人动作自然不生硬，贴近真人。并且该模型在“CCL24-Eval 手语数字人翻译质量评测”任务上取得了综合得分第一的成绩。本文的不足之处在于没有在更细粒度的手语动作特征上进行动作融合，现在无论平滑过渡还是韵律节奏都是以词动作为粒度进行融合，未来为了更贴近真人，应该拆分到左右手、手位置等更细的粒度。同时手语数字人的手语动作不只是局限于手部动作，身体姿态也能传递很多信息，以及如何实现更加生动的面部驱动效果仍是值得研究的方向。

参考文献

Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. 2008. The american sign language lexicon video dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE.

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.

Helen L Bear and Richard Harvey. 2017. Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, 95:40–67.

Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)*, 35(4):1–11.

Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. 2021. Towards fast and high-quality sign language production. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3172–3181.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, December.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. *arXiv preprint arXiv:2010.03142*.

Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 687–705. Springer.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5141–5151.

Ronald W Schafer. 2011. What is a savitzky-golay filter?[lecture notes]. *IEEE Signal processing magazine*, 28(4):111–117.

Kristine Slot. 2007. Motion blending. *Copenhagen University. Department of Computer Science*.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.

Yibo Sun, Duyu Tang, Nan Duan, Yeyun Gong, Xiaocheng Feng, Bing Qin, and Daxin Jiang. 2020. Neural semantic parsing in low-resource settings with back-translation and meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8960–8967.

Guanzhong Tian, Yi Yuan, and Yong Liu. 2019. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks. In *2019 IEEE international conference on Multimedia & Expo Workshops (ICMEW)*, pages 366–371. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ulrich Von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. 2008. The significance of facial features for automatic sign language recognition. In *2008 8th IEEE international conference on automatic face & gesture recognition*, pages 1–6. IEEE.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.