

基于深度神经网络的语音驱动发音器官的运动合成

唐 郅¹ 侯 进¹

摘 要 实现一种基于深度神经网络的语音驱动发音器官运动合成的方法, 并应用于语音驱动虚拟说话人动画合成. 通过深度神经网络 (Deep neural networks, DNN) 学习声学特征与发音器官位置信息之间的映射关系, 系统根据输入的语音数据估计发音器官的运动轨迹, 并将其体现在一个三维虚拟人上面. 首先, 在一系列参数下对比人工神经网络 (Artificial neural network, ANN) 和 DNN 的实验结果, 得到最优网络; 其次, 设置不同上下文声学特征长度并调整隐层单元数, 获取最佳长度; 最后, 选取最优网络结构, 由 DNN 输出的发音器官运动轨迹信息控制发音器官运动合成, 实现虚拟人动画. 实验证明, 本文所实现的动画合成方法高效逼真.

关键词 深度神经网络, 语音驱动, 运动合成, 虚拟说话人

引用格式 唐郅, 侯进. 基于深度神经网络的语音驱动发音器官的运动合成. 自动化学报, 2016, 42(6): 923–930

DOI 10.16383/j.aas.2016.c150726

Speech-driven Articulator Motion Synthesis with Deep Neural Networks

TANG Zhi¹ HOU Jin¹

Abstract This paper implements a deep neural networks (DNN) approach for speech-driven articulator motion synthesis, which is applied to speech-driven talking avatar animation synthesis. We realize acoustic-articulatory mapping by DNN. The input of the system is acoustic speech and the output is the estimated articulatory movements on a three-dimensional avatar. First, through comparison on the performance between ANN and DNN under a series of parameters, the optimal network is obtained. Second, for different context acoustic length configurations, the number of hidden layer units is tuned for best performance. So we get the best context length. Finally, we select the optimal network structure and realize the avatar animation by using the articulatory motion trajectory information output from the DNN to control the articulator motion synthesis. The experiment proves that the method can vividly and efficiently realize talking avatar animation synthesis.

Key words Deep neural networks (DNN), speech-driven, motion synthesis, talking avatar

Citation Tang Zhi, Hou Jin. Speech-driven articulator motion synthesis with deep neural networks. *Acta Automatica Sinica*, 2016, 42(6): 923–930

由于视觉与听觉是人类最主要、最便捷的两种沟通方式, 因此虚拟人动画结合听视觉双模态沟通方式的特点, 将虚拟人的视觉信息作为其声音的一种补充. 例如, 额外的舌头和唇部等发音器官的运动, 眉毛和眼睑等面部特征, 甚至是头部和肢体的动作等, 这些附加信息可以极大提高虚拟人动画的真实感和可懂度. 基于语音驱动虚拟人动画的方法已

经被证实在人机交互应用中十分有效^[1–5].

语音的产生与声道发音器官的运动直接相关, 如唇部、舌头和软腭的位置与移动. 因此, 本文根据声学特征参数估计发音器官的位置信息, 并体现在一个虚拟说话人上面, 实现语音驱动虚拟说话人动画合成. 其中, 最重要的环节是声视觉映射, 即研究声学特征与发音器官位置信息的映射问题.

在最近的十年里, 人工神经网络 (Artificial neural network, ANN)^[6]、隐马尔科夫模型 (Hidden Markov model, HMM)^[7]、高斯混合模型 (Gaussian mixture model, GMM)^[8] 和动态贝叶斯网络 (Dynamic Bayesian network, DBN)^[9] 等被应用于研究声视觉映射问题. 然而, 声学特征与发音器官位置信息之间的映射关系是一个非线性, 多对多的映射问题. 因此, 使用这些算法研究声视觉映射问题的预测精度较低. 在 Uria 等^[10] 和 Zhao 等^[11] 的研究中, 将声学特征与发音器官位置信息之间的映射视为一个回归问题, 使用深度神经网络 (Deep neural

收稿日期 2015-10-31 录用日期 2016-05-03
Manuscript received October 31, 2015; accepted May 3, 2016
成都市科技项目 (科技惠民技术研发项目) (2015-HM01-00050-SF),
四川省动漫研究中心 2015 年度科研项目 (DM201504), 西南交通大学
2015 年研究生创新实验实践项目 (YC201504109) 资助
Supported by Science and Technology Program of Chengdu
(Science and Technology Benefit Project) (2015-HM01-00050-SF),
2015 Annual Research Programs of Sichuan Animation Research
Center (DM201504), and 2015 Graduate Innovative Experimental
Programs of Southwest Jiaotong University (YC2015
04109)
本文责任编辑 柯登峰
Recommended by Associate Editor KE Deng-Feng
1. 西南交通大学信息科学与技术学院 成都 611756
1. School of Information Science and Technology, Southwest
Jiaotong University, Chengdu 611756

networks, DNN) 寻找两者之间的连续映射关系, 并取得良好的实验效果.

在虚拟人面部运动控制问题上, 绝大多数研究者都将发音器官的运动合成作为一个重要的研究方向, 主要体现在唇舌模型的运动控制, 实现虚拟人动画合成. 目前主要有两种主流方法, 一种是基于参数控制的方法^[12-13], 另一种是基于数据驱动的方法^[14-15]. 前者首先建立一个基于二维正面照片的三维人物面部模型, 然后定义一些模型控制参数, 通过计算每一帧动画所需要的参数控制虚拟人面部动画; 后者则是先建立一个图像样本的表情数据库, 在合成阶段根据算法将合适的嘴巴图像从微表情数据库中选出来, 合成情感说话人面部动画.

针对本文实际情况, 采用实验室前期工作, 基于运动轨迹分析的 3D 唇舌肌肉控制模型^[16]. 该模型的优点在于通过分析嘴部和舌部的运动轨迹, 将其分解为一些机械运动的组合, 只需要几个控制参数便能够很好地实现唇部和舌部的自然运动合成.

本文实现一种语音驱动虚拟说话人动画合成方法. 首先, 本文比较基于 ANN 和 DNN 的方法研究声学特征与发音器官位置信息之间映射关系的优劣. 其中, ANN 的网络权值采用随机初始化方式, 而 DNN 采取预训练的方式初始化网络权值. 然后, 在得到较好的网络结构的基础上, 我们进一步研究上下文 (Context) 长度对其重构误差的影响, 获得最佳的 Context 长度. 最后, 在这两个实验结果的基础上, 选取最优网络结构, 由 DNN 输出的发音器官位置信息控制发音器官运动合成, 实现虚拟人动画. 实验证明, 本文所实现的动画合成方法有效逼真.

1 基于深度神经网络的声视觉映射

1.1 深度置信网络

深层次网络训练中的高度非凸性 (Highly non-convex property) 和梯度扩散 (Gradient diffusion) 等问题导致直接训练一个 DNN 是一件很困难的事情. Hinton 等提出一种构建深层次结构神经网络的切实可行的方案^[17]. 该方法的关键在于使用若干个受限的玻尔兹曼机 (Restricted Boltzmann machine, RBM) 无监督生成预训练, 并将这些 RBM 逐层依次向上堆砌成一个 DBN. 生成预训练阶段使每一个 RBM 接近全局最优, 从而确保 DBN 可以获得一个更优的网络权值初值.

1.1.1 受限的玻尔兹曼机

RBM 是一种可以用无向图模型表述的概率模型. 该无向图模型拥有两层结构, 且每一层由若干个概率单元组成. 一个用于描述输入数据特征的可见层 \mathbf{v} 和一个隐藏层 \mathbf{h} . 所有的可见层单元通过一个

无向权值与随机二值的隐藏层单元全连接, 而在可见层和隐藏层的层内单元间无连接.

RBM 是一个基于能量的模型, 在模型参数 θ 下, 记其可见层和隐藏层的联合组态为 $(\mathbf{v}, \mathbf{h}; \theta)$, 其能量函数为 $E(\mathbf{v}, \mathbf{h}; \theta)$. 则可见层 \mathbf{v} 与隐藏层 \mathbf{h} 的联合概率分布为

$$p(\mathbf{v}, \mathbf{h}; \theta) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h}; \theta)) \quad (1)$$

其中, $\theta = \{W, \mathbf{b}, \mathbf{c}\}$. W 是可见层与隐藏层之间的连接权值矩阵; \mathbf{b}, \mathbf{c} 分别是可见层和隐藏层的偏置向量; $Z = \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \theta))$, 为配分函数 (Partition function).

当 RBM 的可见层和隐藏层单元都是随机二值类型时, 我们采用 Bernoulli-Bernoulli RBM (二值 RBM). 其联合概率分布的能量函数被定义为

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\mathbf{v}^T W \mathbf{h} - \mathbf{b}^T \mathbf{v} - \mathbf{c}^T \mathbf{h} \quad (2)$$

当可见层输入是实际的特征值时, 如语音参数梅尔倒谱系数 (Mel-scale frequency cepstral coefficients, MFCC), 而隐藏层是随机二值类型时, 我们采用 Gaussian-Bernoulli RBM (GRBM)^[18].

通常, 我们将 GRBM 输入端的实际的特征数据进行归一化处理, 使其具有 0 均值且标准差为 1. 则其联合概率分布的能量函数被定义为

$$E(\mathbf{v}, \mathbf{h}; \theta) = \frac{1}{2}(\mathbf{v} - \mathbf{b})^T (\mathbf{v} - \mathbf{b}) - \mathbf{v}^T W \mathbf{h} - \mathbf{c}^T \mathbf{h} \quad (3)$$

在 RBM 的生成训练阶段, 我们使用对比散度 (Contrastive divergence, CD)^[19] 算法.

1.1.2 堆砌 RBM 成 DBN

我们将若干个 RBM 自下而上一层一层地堆砌成 DBN, 堆砌规则可参见文献 [20]. 因为本文中输入的数据为声学特征, 故最低层 RBM 本文采用 GRBM, 其他层为二值 RBM. 将 GRBM 隐藏层单元的状态作为新数据, 用于训练更高一层的二值 RBM; 而在两个二值 RBM 之间使用低层的输出值作为更高一层的输入数据. 采用这种重复的方法, 我们可以获得期望的隐藏层层数的网络结构.

1.2 搭建并微调 DNN 结构

本文在 DBN 的最顶层增加一个线性输出层形成 DNN^[21], 用于研究声学特征与发音器官位置信息之间的映射问题. 输入为语音特征参数, 输出为发音器官的位置信息. 使用预训练 DBN 获得的各层参数依次初始化与 DNN 对应的每一层, 这样我们便可以获得一个接近最优参数的深层网络结构. 最后, 我们便可以将 DNN 当作传统的 ANN, 使用误差反向传播 (Error back propagation, BP) 算法进行微调网络参数.

2 语料库

本文使用 MNGU0 数据库^[22] 研究声学特征与发音器官位置信息之间的映射问题. 该数据库采用电磁关节造影技术 (Electromagnetic articulography, EMA) 并行记录一个说话者说话时发音器官的位置信息, 同时记录说话者的语音数据资料. 如图 1 所示, 分别记录上唇 (UL)、下唇 (LL)、下颌切牙 (LI)、舌尖 (T1)、舌片 (T2) 和舌背 (T3) 上观测点的位置信息. EMA 以 200 Hz 的采样频率记录这 6 个观测点的 x 和 y 轴坐标值, 共计 12 维数据. 至于音频数据, 首先将记录的语音数据降低采样频率至 16 kHz, 然后使用 STRAIGHT^[23] 提取 40 维频率扭曲线谱频率 (Frequency-warped line spectral frequencies, LSFs), 并加一个增益值. 在所有的 EMA 和 LSFs 参数向量的每一个维度上, 先减去其平均值, 然后除以 4 倍的标准差, 进行归一化处理.

MNGU0 数据库包含 1354 个语音片段文件和对应的 EMA 数据文件. 其中, 校验和测试数据集各具有 63 个音频和对应的 EMA 数据文件, 则剩余的 1228 个音频和对应的 EMA 数据文件作为训练数据集.

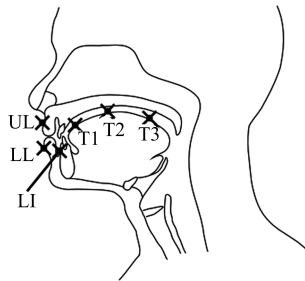


图 1 MNGU0 数据库中 EMA 记录发音器官的 6 个观测点^[22]

Fig.1 Positioning of the six electromagnetic coils in the MNGU0 dataset^[22]

3 发音器官模型

本文主要驱动发音器官为嘴部和舌部, 我们采用实验室前期工作基于运动轨迹分析的 3D 唇舌肌肉控制模型^[16]. 图 2 和图 3 分别代表三维虚拟人唇部模型和舌部模型. 该模型的优点在于通过分析嘴部和舌部的运动轨迹, 将其分解为一些机械运动的组合, 只需通过计算口轮匝肌外圈肌、舌纵肌等的肌肉收缩量 oos 、 zt 和下颌的旋转角度 jaw , 便可以很好地实现唇部和舌部的自然运动合成.

3.1 计算控制量 oos 、 zt 和 jaw

根据文献 [16], 推出口轮匝肌外圈肌的肌肉收缩量的计算公式如下:

$$oos = \frac{\Delta x h_m}{K_x h_t \cos\left(\frac{h_m}{L_s}\right)} \quad (4)$$

其中, Δx 为预测出上唇的 x 坐标相对其初始状态的相对变化量; h_m 为初始状态下唇舌模型的上下嘴唇高度差; h_t 为初始状态下测量的上唇与下唇的 y 坐标的相对差值; K_x 为伸缩系数, 通过实验获得 $K_x = 0.2$; L_s 为唇舌模型中唇部长度.

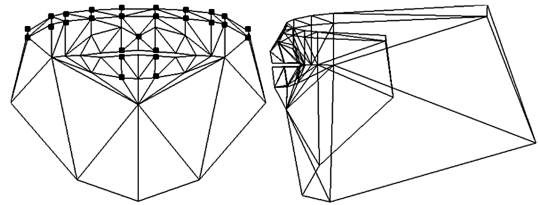


图 2 嘴部网格模型

Fig.2 Mouth mesh model

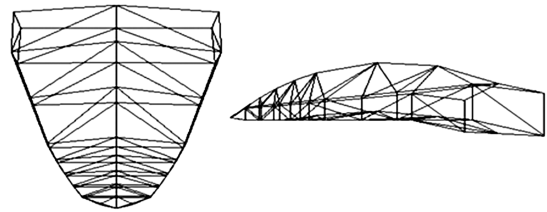


图 3 舌部网格模型

Fig.3 Tongue mesh model

计算下颌的旋转角度 jaw 的方法如图 4 所示. 其中, 点 U 和 L 分别表示上唇和下唇的测量点位置; O' 为线段 UL 的中点; 点 J 为初始状态下下颌切牙测量点位置; J' 为说话时下颌切牙的一个位置. 则计算下颌的旋转角度公式为

$$jaw = \tan^{-1}\left(\frac{QJ'}{QO'}\right) - \tan^{-1}\left(\frac{PJ}{PO'}\right) \quad (5)$$

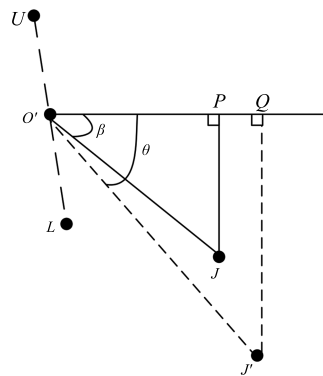


图 4 下颌的旋转角度分析

Fig.4 The rotation of the mandible angle analysis

根据文献 [16], 推出舌纵肌的肌肉收缩量的计算公式如下:

$$zt = K_d d_m \frac{d_{T'} - d_T}{d_T} \quad (6)$$

其中, d_T 为初始状态下舌片 T2 与舌尖 T1 之间的距离; $d_{T'}$ 为说话时舌片 T2 与舌尖 T1 之间的距离; K_d 为伸缩系数, 通过实验获得 $K_d = 0.05$; d_m 为唇舌模型中舌片与舌尖之间的距离.

我们通过 DNN 预测输出的发音器官位置信息可以计算出轮匝肌外圈肌、舌纵肌等的肌肉收缩量 oos 、 zt 和下颌的旋转角度 jaw , 从而实现发音器官的运动合成.

4 实验结果与分析

本文的实验环境为 Intel Xeon E3-1231 v3 3.4 GHz, 16 GB 内存, Window 7, Matlab 2012b, VS2010, OpenGL.

采用均方根误差 (Root mean-squared error, RMSE) 评价基于神经网络的方法实现声学特征与发音器官位置信息之间映射关系的实验效果, 其定义如下:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (e_i - t_i)^2} \quad (7)$$

其中, e_i 为发音器官位置信息的估计值; t_i 为发音器官位置信息的真实测量值.

本文分别计算发音器官位置信息的每一维 RMSE, 然后取 12 维 RMSE 的平均值作为最终的重构误差 RMSE.

采用文献 [24] 的客观评价方法评价唇部动画合成的真实度. 本文考查实际唇部的归一化高度与动画合成唇部的归一化高度的差值 Δh :

$$\Delta h = \left| \frac{h_{\text{real}}}{h_{r0}} - \frac{h_{\text{syn}}}{h_{s0}} \right| \quad (8)$$

其中, h_{real} 与 h_{syn} 分别是在发音阶段实际唇部和动画合成唇部的高度; h_{r0} 与 h_{s0} 分别是在唇部自然闭合状态下实际唇部和动画合成唇部的高度.

计算 Δh 的均值和方差, 其均值 E 与方差 Var 越小, 则合成口型与实际口型越吻合. 按照 5 分制建立客观评测值 Obj :

$$Obj = \frac{5}{1 + E + \text{Var}} \quad (9)$$

本文分别计算每个测试数据的客观评测值 Obj , 并取其均值, 作为最终的动画合成评测结果.

4.1 实验条件

本文使用 N 个语音帧组成的上下文窗 (Context window) 作为 DNN 的输入数据, 调整 N 获得最优的实验结果. 因为本文使用 41 维语音参数 (40 维 LSFs 加 1 个增益值), 因此 DNN 的输入层单元数为 $41 \times N$. 至于输出端, 本文不仅选择与当前上下文窗的中间时刻对应的 12 维 EMA 数据, 而且还考虑 EMA 数据的一阶和二阶差分. 所以, DNN 输出层含有 36 个单元.

在本文中, 输入层为 GRBM, 其他层为二值 RBM. 在预训练阶段, 本文设置所有 RBM 的小批量 (Mini-batch) 为 128, 动量因子为 0.9, 未使用权重衰减. 设置 GRBM 的学习速率为 0.001, 迭代 50 次; 而二值 RBM 的学习速率为 0.01, 迭代 10 次.

在 DNN 调整网络权值阶段, 本文采用 BP 算法的随机梯度下降法微调网络权值, 且小批量同为 128. 设置网络的学习速率为 0.01, 动量因子为 0.9, 迭代 500 次. 在每次迭代时, 网络学习速率的衰减因子为 0.99.

4.2 ANN 和 DNN 实验结果的对比

本文进行一系列实验, 比较 ANN 和 DNN 的实验结果优劣. 在此次实验中, 本文采用 10 个语音帧组成的上下文窗作为 DNN 的输入层, 故输入层有 410 个节点单元. ANN 和 DNN 网络分别含有 1 至 4 个隐藏层, 且每个隐藏层分别有 100、200、300 和 400 个单元, 均采用测试数据集进行测试, 共得到 32 组实验结果, 如图 5 所示.

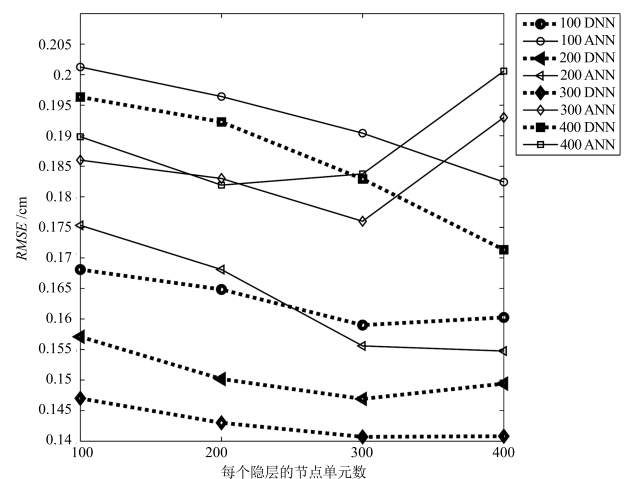


图 5 对比 ANN 和 DNN 的实验结果

Fig. 5 Comparison on the experimental results of ANN and DNN

从图 5 中可知, 基于 DNN 方法的重构误差明显小于 ANN. 因此, 使用基于 DNN 的方法研究声学特征与发音器官位置信息之间映射关系的效果更

优. 另外, 当 DNN 含有 3 个隐藏层, 且每个隐藏层内具有 300 个单元时, 在 10 个语音帧组成的上下文窗的条件下, 其重构误差最小.

文献 [11] 中也选择 MNGU0 数据库作为训练与测试, 其 DNN 含有 4 个隐藏层, 且每个隐藏层均

含有 1000 个单元的网络结构. 从文献 [11] 的结果图中可以看出, 其最优结果大于 0.145 cm, 而本文所得到的最小重构误差小于文献 [11] 中的结果, 效果更优.

图 6 为由 400 帧语音参数估计出的舌尖 (T1)、

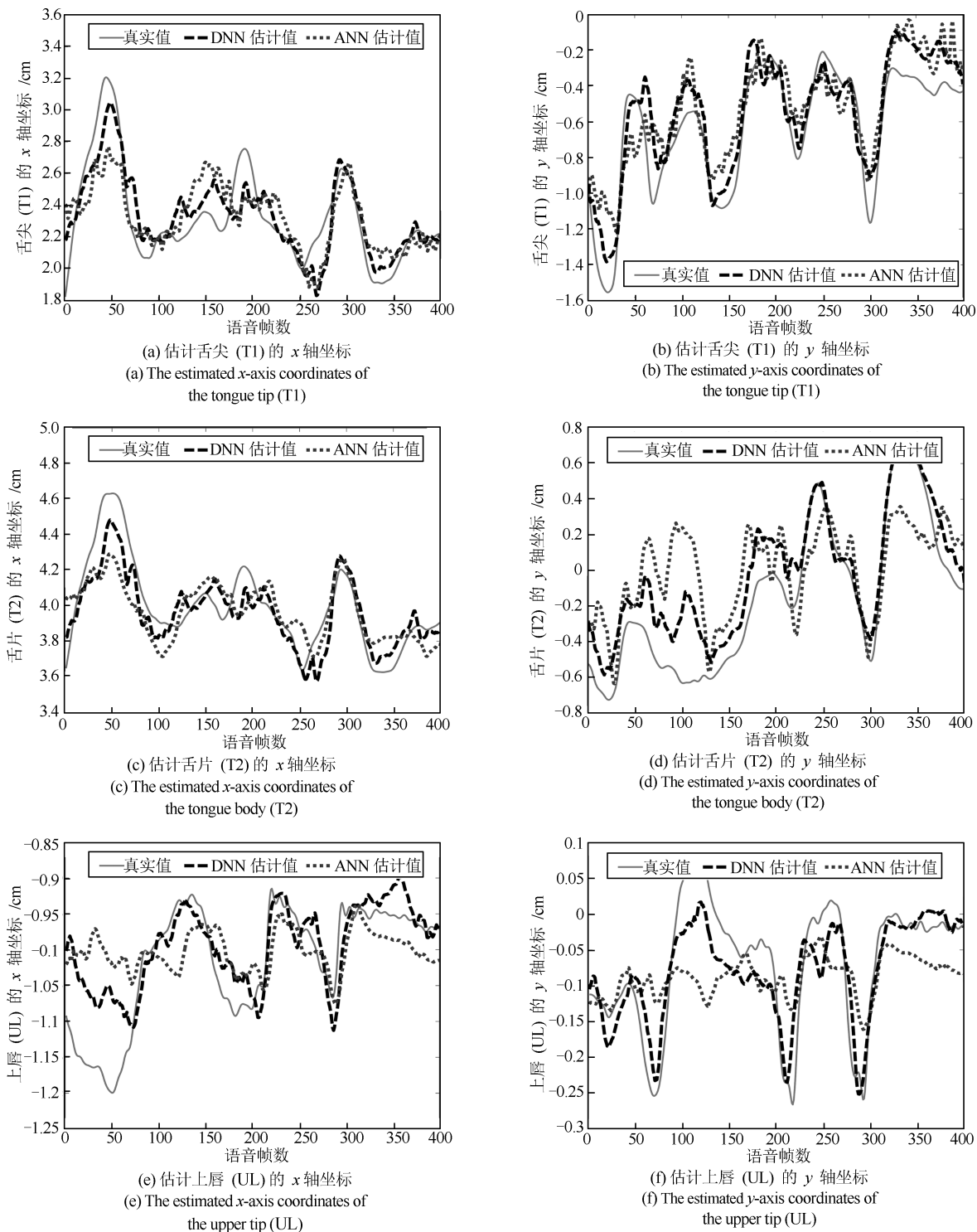


图 6 比较 ANN 和 DNN 估计的发音器官运动轨迹

Fig. 6 Comparison on the estimated articulatory motion trajectories between ANN and DNN

舌片 (T2) 和上唇 (UL) 运动轨迹. 图中实线为真实值, 点虚线为 ANN 估计值, 短线虚线为 DNN 估计值. 从图 6 中可以看出, 采用基于 DNN 的方法拟合出特征点的运动轨迹更接近真实的运动轨迹, 故 DNN 的拟合效果优于 ANN. 通过对比重构误差和拟合 T2 的运动轨迹这两个方面, 均可以发现 DNN 优于 ANN.

4.3 上下文的长度对实验结果的影响

在基于深度学习的语音识别领域, 研究人员发现将长的上下文声学特征作为输入端, 可以获得更好的识别效果^[25]. 因此, 本文尝试寻找最佳的上下文窗作为 DNN 的输入数据, 使估计出发音器官的位置信息重构误差最小. 本文试验 6~30 个语音帧组成的上下文窗, 且每次增加 4 帧. 我们采用含有 3 个隐藏层的预训练网络, 且在训练阶段均迭代 500 次, 调整隐藏层层内单元数使网络最优.

实验结果如表 1 所示, 从中我们发现适当地增加上下文窗的长度可以降低重构误差, 但过长的上下文窗并不能取得更好效果. 因此, 使用适当长度的上下文声学特征作为 DNN 的输入端, 可以有效地降低重构误差. 本文上下文长度为 22 帧, 且 DNN 每层含有 500 隐藏单元时, RMSE 最小, 为 0.134 cm.

表 1 上下文窗的长度对 RMSE 的影响

Table 1 Effect of the length of the context window on the RMSE

上下文长度 (帧数)	RMSE (cm)
6	0.149
10	0.141
14	0.138
18	0.135
22	0.134
26	0.134
30	0.136

4.4 唇部动画评价结果

本文在前两个实验结果的基础上, 选取含有 3 个隐藏层, 每层含有 500 个单元的深度学习神经网络, 其 DNN 输入端为 22 帧语音参数的最优网络. 我们使用得到的 12 维发音器官位置信息控制虚拟人动画合成.

通过本文使用的方法合成测试集中的一段 House shook 的三维人脸口型动画. 为了方便而直观地验证人脸口型动画的逼真度, 录制真实人脸视频与基于本文方法合成的动画进行对比, 如图 7 所示.

图 7 中第 1 行为真实人脸在说话时的口型截图, 第 2 行为本文方法合成的三维人脸口型动画截图. 通过主观对比评价可以发现, 基于本文方法合成的动画与真实说话人发音口型变化规律相同.

采用客观评价方法对本文合成的 63 个测试动画进行客观评测, 得到评测结果如表 2 所示. 结果表明基于本文方法实现的动画效果比传统方法较优, 并且动画合成更加简易.

表 2 客观评价结果

Table 2 Objective assessment results

	传统方法	本文方法
Obj	3.6	3.7

通过主客观评价分析, 得出本文所实现的方法接近真实说话人的口型动画变化趋势, 并且合成的口型动画的综合客观评价结果较好. 因此, 实验结果证明本文所实现的动画合成方法简易有效且逼真.

5 结论

本文实现一种基于深度神经网络的语音驱动发音器官运动合成的方法, 将其用于语音驱动虚拟说话人系统. 通过 DNN 学习语音特征与发音器官位置

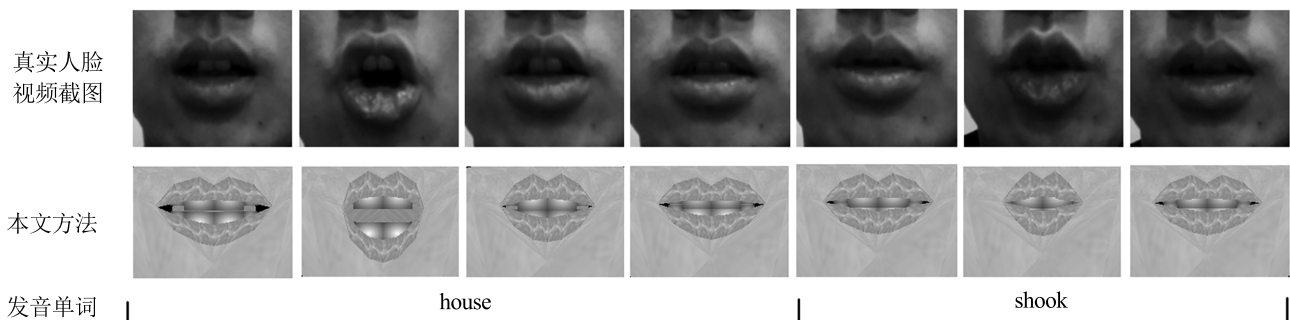


图 7 口型动画部分截图

Fig. 7 Snapshots from the lip animation

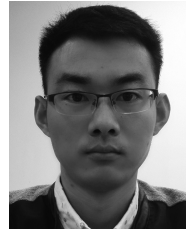
信息之间的映射关系, 从而根据输入的语音数据估计出发音器官的运动轨迹, 并将其体现在一个三维虚拟人上面. 首先, 本文在一系列参数下对比 ANN 和 DNN 的实验结果, 得到最优网络; 其次, 设置不同上下文声学特征长度并调整隐层单元数, 获取最佳长度; 最后, 本文在这两个结论的基础上, 选取最优网络结构, 由 DNN 输出的发音器官位置信息控制发音器官运动合成, 实现虚拟人动画. 实验证明, 本文所实现的动画合成方法高效且逼真, 优点在于合成动画的控制参数少, 简单方便. 但是, 也存在一些问题, 如本文只使用了舌纵肌控制舌部动画合成, 实现舌头卷起动作, 而未考虑控制舌头厚度变化. 因此, 在未来工作中会对其进行改善, 生成更加逼真的舌部动画.

References

- Liu J, You M Y, Chen C, Song M L. Real-time speech-driven animation of expressive talking faces. *International Journal of General Systems*, 2011, **40**(4): 439–455
- Le B H, Ma X H, Deng Z G. Live speech driven head-and-eye motion generators. *IEEE Transactions on Visualization and Computer Graphics*, 2012, **18**(11): 1902–1914
- Han W, Wang L J, Soong F, Yuan B. Improved minimum converted trajectory error training for real-time speech-to-lips conversion. In: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Kyoto, Japan: IEEE, 2012. 4513–4516
- Ben-Youssef A, Shimodaira H, Braude D A. Speech driven talking head from estimated articulatory features. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014. 4573–4577
- Ding C, Zhu P C, Xie L, Jiang D M, Fu Z H. Speech-driven head motion synthesis using neural networks. In: Proceedings of the 2014 Annual Conference of the International Speech Communication Association (INTER-SPEECH). Singapore, Singapore: ISCA, 2014. 2303–2307
- Richmond K, King S, Taylor P. Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech and Language*, 2003, **17**(2–3): 153–172
- Zhang L, Renals S. Acoustic-articulatory modeling with the trajectory HMM. *IEEE Signal Processing Letters*, 2008, **15**: 245–248
- Toda T, Black A W, Tokuda K. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, 2008, **50**(3): 215–227
- Xie L, Liu Z Q. Realistic mouth-synching for speech-driven talking face using articulatory modelling. *IEEE Transactions on Multimedia*, 2007, **9**(3): 500–510
- Uria B, Renals S, Richmond K. A deep neural network for acoustic-articulatory speech inversion. In: Proceedings of the 2011 NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Granada, Spain: NIPS, 2011. 1–9
- Zhao K, Wu Z Y, Cai L H. A real-time speech driven talking avatar based on deep neural network. In: Proceedings of the 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Kaohsiung, China: IEEE, 2013. 1–4
- Tang H, Fu Y, Tu J L, Hasegawa J M, Huang T S. Humanoid audio-visual avatar with emotive text-to-speech synthesis. *IEEE Transactions on Multimedia*, 2008, **10**(6): 969–981
- Fu Y, Li R X, Huang T S, Danielsen M. Real-time multimodal human-avatar interaction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2008, **18**(4): 467–477
- Schreer O, Englert R, Eisert P, Tanger R. Real-time vision and speech driven avatars for multimedia applications. *IEEE Transactions on Multimedia*, 2008, **10**(3): 352–360
- Liu K, Ostermann J. Realistic facial expression synthesis for an image-based talking head. In: Proceedings of the 2011 IEEE International Conference on Multimedia and Expo (ICME). Barcelona, Spain: IEEE, 2011. 1–6
- Yang Yi, Hou Jin, Wang Xian. Mouth and tongue model controlled by muscles based on motion trail analyzing. *Application Research of Computers*, 2013, **30**(7): 2236–2240 (杨逸, 侯进, 王献. 基于运动轨迹分析的 3D 唇舌肌肉控制模型. 计算机应用研究, 2013, **30**(7): 2236–2240)
- Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504–507
- Hinton G E. A practical guide to training restricted Boltzmann machines. *Neural Networks: Tricks of the Trade* (2nd Edition). Berlin: Springer-Verlag, 2012. 599–619
- Tieleman T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In: Proceedings of the 25th International Conference on Machine Learning (ICML). New York, USA: ACM, 2008. 1064–1071
- Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527–1554
- Hinton G, Deng L, Yu D, Dahl G E, Mohamed A R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T N, Kingsbury

- B. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, 2012, **29**(6): 82–97
- 22 Richmond K, Hoole P, King S. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory. In: Proceedings of the 2001 Annual Conference of the International Speech Communication Association (INTER-SPEECH). Florence, Italy: ISCA, 2011. 1505–1508
- 23 Kawahara H, Estill J, Fujimura O. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In: Proceedings of the 2nd International Workshop Models and Analysis of Vocal Emissions for Biomedical Application (MAVEBA). Firenze, Italy, 2001. 59–64
- 24 Li Hao, Chen Yan-Yan, Tang Chao-Jing. Dynamic Chinese visemes implemented by lip sub-movements and weighting function. *Signal Processing*, 2012, **28**(3): 322–328
(李皓, 陈艳艳, 唐朝京. 唇部子运动与权重函数表征的汉语动态视位. *信号处理*, 2012, **28**(3): 322–328)
- 25 Deng L, Li J Y, Huang J T, Yao K S, Yu D, Seide F, Seltzer M, Zweig G, He X D, Williams J, Gong Y F, Acero A. Recent advances in deep learning for speech research at Microsoft.

In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, Canada: IEEE, 2013. 8604–8608



唐 郅 西南交通大学信息科学与技术学院硕士研究生. 主要研究方向为虚拟说话人动画与模式识别.

E-mail: tang_zhi@126.com

(**TANG Zhi** Master student at the School of Information Science and Technology, Southwest Jiaotong University. His research interest covers talking avatar animation and pattern recognition.)



侯 进 西南交通大学信息科学与技术学院副教授. 主要研究方向为计算机动画, 数字艺术和自动驾驶. 本文通信作者. E-mail: jhou@swjtu.edu.cn

(**HOU Jin** Associate professor at the School of Information Science and Technology, Southwest Jiaotong University. Her research interest covers computer animation, digital art, and automatic driving. Corresponding author of this paper.)