ISSN 1000-0054 清华大学学报（自然科学版）2013 年 第 53 卷 第 6 期 35/40
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2013, Vol. 53, No. 6 898-902

# 语音驱动虚拟说话人的自然头动生成

李冰锋， 谢 磊， 朱鹏程， 樊 博

（西北工业大学 计算机学院，陕西省语音与图像信息处理重点实验室，西安 710072）

**摘 要**：从语音信号预测伴随头动时，基于隐 Markov 模型（hidden Markov model，HMM）的头动合成方法的效果依赖于头动模式的划分和头动模式的正确识别。该文尝试了不同头动模式划分方法的头动合成效果。由于语音和头动之间是非确定性的多对多的映射关系，很难用固定的类别描述清楚，因此该类方法的头动模式识别率不高，头动合成效果受限。该文尝试采用逆传播（back-propagation，BP）神经网络的非线性回归方法，通过学习语音与头动之间的映射关系，实现语音信号到头动参数之间的直接连续映射，避免了 HMM 方法中头动模式不明确、头动模式识别错误带来的负面影响。实验表明，基于 BP 神经网络的回归方法有效地提高了语音到头动预测的准确度和头动合成的自然度。

**关键词**：虚拟说话人；面部动画；头动生成；隐 Markov 模型；神经网络

**中图分类号**：TP 391 **文献标志码**：A

**文章编号**：1000-0054(2013)06-0898-05

## Head motion generation for speech-driven talking avatar

LI Bingfeng, XIE Lei, ZHU Pengcheng, FAN Bo

(Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

**Abstract**: This study describes methods for predicting head motion from acoustic speech. Current hidden Markov model (HMM)-based methods rely on definitions of typical head motion patterns and accurate recognition of these patterns. This study investigates the head motion prediction performance of various pattern definition strategies. The HMM method is less effective because the association between speech and the head gestures is essentially a nondeterministic, many-to-many mapping so the head motion pattern recognition accuracy is quite low. Therefore, this study treats the speech-to-head-motion mapping task as a regression problem. A back-propagation (BP) neutral network is used to seek a direct, continuous mapping from the acoustic speech to the head motion. Tests show that this neutral network approach significantly improves the head motion prediction accuracy and the naturalness of head movement of a talking avatar.

**Key words**: talking avatar; talking head; head motion generation; hidden Markov model; neural network

虚拟说话人（talking avatar）是指由计算机生成的能够开口说话的虚拟人物形象[1]。人们在进行语音交流的时候，总伴随着非言语的头部动作，例如点头、眨眼等。这些自然的动作具有重要的辅助言语的提示性作用，也反映了说话人的情感状态。因此，实现逼真的虚拟说话人，不仅需要同步一致的口型动作，还需要通过头部运动、面部表情来传递表达丰富的非言语信息。自然的头部运动在语音交流过程中十分重要。研究表明，头动和语音具有明显的相关性[2]。Munhall 发现自然头动可以明显提高语音可懂度[3]。McNeill 认为头动是语音产生过程的自然产物，属于语音生成的一部分[4]。因此，自然头动是提高虚拟说话人表现力的关键因素。

根据输入信息的不同，头动生成可以分为基于文本和基于语音的方法。基于文本的方法通过分析头部运动与语言文本韵律结构、语义信息之间的相互关系，建立头部运动和文本语义标注之间的对应规则[5]或关联模型[6]，进而实现以文本韵律词为合成单位的头部运动生成算法。既然语音和头动之间存在着明显的相关性，头动是语音产生过程的产物，因此从语音的角度出发研究头动也是一条可取的途径。此类方法需要录制说话人讲话的音视频数据，提取能量、基频等语音韵律特征，探寻语音特征与伴随头动之间的映射关系，建立音视频关联模型，在模型的基础上从语音输入中预测头动轨迹。

**收稿日期**：2013-04-27

**基金项目**：国家自然科学基金面上项目（61175018）；
陕西省自然科学基础研究计划（2011JM8009）；
霍英东基金项目（131059）

**作者简介**：李冰锋（1988—），男（汉），河南。

**通信作者**：谢磊，教授，E-mail：lxie@nwpu.edu.cn

近年来，隐 Markov 模型(hidden Markov model，HMM)[7-10]和混合 Gauss 模型(Gaussian mixture model，GMM)[11]被广泛应用于语音驱动的头动生成。以基于 HMM 的方法为例，该类方法首先确定头动的模式或类别，然后使用提取的语音特征和对应的头动特征，为每个类别训练一个“语音-头动”双模态 HMM 模型，建立语音与头动之间的映射关系。头动的类别可以人工指定，或者采用聚类方法自动确定。例如，Graf 等通过统计实验将头部运动分为 3 个模式：“V 型”、“∩型”和“/型”[12]。Busso 等[7]采用矢量量化方法，将头部的连续运动轨迹自动划分为若干离散的头部姿态。在头动合成阶段，根据“语音-头动”HMM 模型，将输入语音识别成头动模型序列，串接模型对应的头动 HMM 参数生成头动序列。最直接的方法是采用 HMM 模型各个状态对应头动参数的 Gauss 均值作为输出头动。然而，这种方法会使头部出现明显的跳动，这是因为在模型衔接处存在参数不连续的现象。此外，由于仅使用均值，不考虑方差，头动幅度受限，与真实头动差距较大。为解决这一问题，Busso 等将从模型的协方差矩阵中产生的彩色噪声，加入到均值序列中，同时对衔接处进行差值[7]。最近，Hofer 等从基于 HMM 的语音合成[13]中得到启发，采用 Trajectory HMM 算法，在极大似然准则下，利用动态特征生成平滑的头动曲线，提高了头动合成的自然度[9-10]。

然而，这种基于头动模式“识别”的头动合成方法存在严重问题。首先，识别率不高，识别错误严重影响头动合成的效果。这主要是因为语音和头动之间是非确定性的多对多的映射关系[12]；且不同说话人、不同语音内容的头动模式也存在较大差别。头动仅仅是角度 3 个自由度的连续旋转，而语音信号中承载着复杂的言语信息。研究指出，头动与语音能量的波峰具有明显的相关性，然而波峰的级别与不同头动类型之间没有明确的相关性[14]。

针对这些问题，本文尝试了不同头动模式划分方法在基于 Trajectory HMM 头动合成上的效果，以期找出最优方案。同时，尝试采用回归方法学习语音与头动之间非确定多对多的连续映射关系。回归方法通过学习从输入变量到输出变量之间的映射函数，预测输入变量和输出变量之间的关系。本文训练逆传播(back propagation，BP)神经网络，实现语音信号到头动参数之间的直接连续映射，避免了头动模式识别方法中“头动模式不明确”、识别错误、参数插值平滑带来的负面影响。实验表明，基于 BP 神经网络的回归方法明显优于基于 Trajectory HMM 方法，有效地提高了语音到头动预测的准确度和头动合成的自然度。

## 1 数据准备与特征提取

本文采用主持人新闻播报作为研究对象，从 CCTV《新闻直播间》节目中手工切出了一个主持人播报新闻的音视频数据，将其切分为 163 句，每句时长 10～15 s 左右。本文采用 AAM-FPT 面部跟踪工具[15]提取头部相对于 $X$，$Y$，$Z$ 轴转动的 Euler 角度，这 3 个角度分别对应了点头(Pitch)、摇头(Yaw)和摆头(Roll)。图 1 展示了一个视频对应的头动 Euler 角度曲线。由于视频帧率是 25 帧/s，因此每 40 ms 得到一次头动角度。可以看出，伴随着讲话，主持人具有明显的头动。为了能够反映头部运动的动态变化，本文使用这 3 个头动角度以及它们的一、二阶差分作为头动特征向量，共计 9 维。根据前期研究结果[11]，对应的音频特征选用平滑后的 Mel 频率倒谱系数(Mel frequency ceptral coefficient，MFCC)(Smoothed MFCC)、短时能量(Energy)和基频(Pitch)，以及它们的一阶、二阶差分，共得到 42 维音频特征。特征提取采用 HTK 工具包，窗口长度设为 25 ms，帧移为 15 ms。

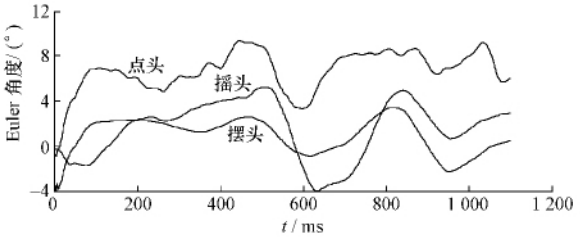


图 1 一个新闻视频中主持人的头动 Euler 角度曲线

## 2 基于 Trajectory HMM 的头动合成

从基于 HMM 的语音合成中得到启发，Hofer 等采用 Trajectory HMM 算法得到了连续的头动曲线[10]，该系统框架如图 2 所示。该方案将头部运动姿态划分为几个典型的类别，提取音视频特征，为每一个类别训练一个音视频 HMM 模型。在头动合成阶段，根据 HMM 模型的音频部分，将输入语音首先识别成头动类别序列，然后采用基于极大似然准则(maximum likelihood，ML)的曲线生成算法，通过动态参数的(一、二阶差分)引入，输出连续光滑的头动曲线[10,13]。

基于 HMM 的头动参数合成算法，需要首先将

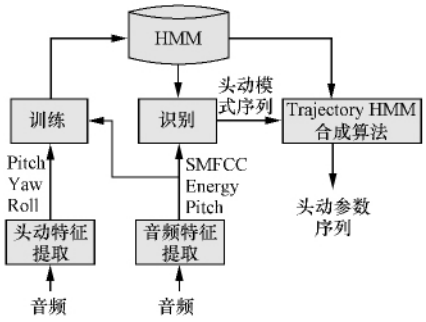


图 2 基于 Trajectory HMM 的头动合成系统框架[10]

头部运动划分成不同的类别。划分方法主要有两类：1）根据观察到的头部重复出现的典型模式进行人工划分；2）使用聚类算法，按照头部姿态在 3 维空间中的位置进行自动划分。

本文具体的分类方法如下：

1）手工标注 4 个类别。Graf 等通过统计实验，将头部运动分为 3 个模式[12]：(a)“V 型”：头部快速运动并迅速回复到原始位置，运动轨迹类似 V 字型；(b)“∩型”：头部快速运动并迅速回复且超过原始位置；(c)“/型”：头部快速运动，但不再回复到原位置。(d) 本文将头部没有明显运动的情况归结为一个头部相对静止类，命名为“P 型”。

2）手工标注 6 个类别。本文通过观察数据库中受测者的头动视频，将 Garf 等定义的 3 个类别以及“P 型”进行了细化，得到了 6 个精细类别。(a)“ZZ 型”：头部端正，没有明显的动作；(b)“UZ 型”：只有点头的动作，点头之后快速回正(或者不回正)；(c)“ZU 型”：没有点头，向左(或者右)快速摆头，然后回正(或者不回正)；(d)“DZ 型”：头部从正常状态，向后仰；(e)“UU 型”：既有点头，又有摇头，且动作幅度较大；(f)“PA 型”：除上述 5 种类别之外的动作。

3）基于 $K$-means 的聚类。为了能够自动标注抄本，本文采用 $K$-means 聚类方法，对头动特征进行自动聚类，这种聚类方法实质是按照头部在 3 维空间的位置进行划分，每一类代表了头部在空间中的一个姿态。经过经验实验对比，本文最终将头部运动聚为了 8 类。这种聚类方法具有较好的推广性，对于不同数据库、不同的受测者均适用。

## 3 基于神经网络的头动合成

上述 HMM 头动合成方法有赖于头动单元的正确识别。然而，识别错误不可避免，这些错误严重影响头动合成的效果。识别效果不佳的原因在于语音和头动之间是非确定性的多对多映射关系[12]。因此，本文尝试训练 BP 神经网络，将语音到头动的合成问题转换为回归问题，通过学习获得语音信号到头动参数之间的直接映射关系。该方法不需要将头部运动划分成不同的类别，直接使用提取得到的音频特征和头动特征训练神经网络的各个参数，以实现音频特征到头动特征的直接映射。

BP 神经网络模型采用误差反向传播算法，近似实现从输入到输出的任意连续的非线性映射[16]。由于输入输出特征的维数并不高，因此本文选用了 3 层网络结构，如图 3 所示。假设输入音频特征向量为 $\boldsymbol{o}_{\mathrm{S}}$，维数为 $D_{\mathrm{S}}$，经过 BP 网络拟合输出的头动特征向量为 $\boldsymbol{o}_{\mathrm{H}}$，维数为 $D_{\mathrm{H}}$，网络中间隐层的个数为 $N$，则输入输出的直接映射关系由式(1)—(5)实现：

$$\boldsymbol{o}_{\mathrm{H}} = \boldsymbol{W}^{\mathrm{out}}(\mathrm{tansig}(\boldsymbol{W}^{\mathrm{in}}\boldsymbol{o}_{\mathrm{S}} + \boldsymbol{b}^{\mathrm{in}})) + \boldsymbol{b}^{\mathrm{out}}), \quad (1)$$

$$\boldsymbol{W}^{\mathrm{in}} = \begin{bmatrix} w_{11}^{\mathrm{in}} & \cdots & w_{1D_{\mathrm{S}}}^{\mathrm{in}} \\ \vdots & \ddots & \vdots \\ w_{N1}^{\mathrm{in}} & \cdots & w_{ND_{\mathrm{S}}}^{\mathrm{in}} \end{bmatrix}, \quad (2)$$

$$\boldsymbol{b}^{\mathrm{in}} = [b_1^{\mathrm{in}}, \cdots, b_N^{\mathrm{in}}]^{\mathrm{T}}, \quad (3)$$

$$\boldsymbol{W}^{\mathrm{out}} = \begin{bmatrix} w_{11}^{\mathrm{out}} & \cdots & w_{1N}^{\mathrm{out}} \\ \vdots & \ddots & \vdots \\ w_{D_{\mathrm{H}}1}^{\mathrm{out}} & \cdots & w_{D_{\mathrm{H}}N}^{\mathrm{out}} \end{bmatrix}, \quad (4)$$

$$\boldsymbol{b}^{\mathrm{out}} = [b_1^{\mathrm{out}}, \cdots, b_{D_{\mathrm{H}}}^{\mathrm{out}}]^{\mathrm{T}}. \quad (5)$$

其中：隐层传递函数采用非线性的双曲正切 S 型传递函数 tansig，输出层采用线性函数 purelin。$\boldsymbol{W}^{\mathrm{in}}$ 和 $\boldsymbol{W}^{\mathrm{out}}$ 分别为输入层和输出层的权值矩阵，$\boldsymbol{b}^{\mathrm{in}}$ 和 $\boldsymbol{b}^{\mathrm{out}}$ 分别为输入层和输出层的偏差因子。

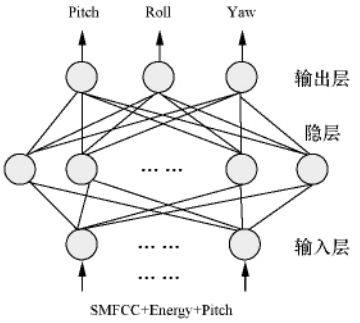


图 3 神经网络结构示意图

## 4 实 验

### 4.1 实验设置

本文从数据库 163 句话中随机抽取了 133 句作为训练集，剩余 30 句作为测试集。本文对 5 种头动合成方案进行了对比实验：基于 Trajectory-HMM

手工标注 4 类、手工标注 6 类、自动标注聚 8 类的方法，基于 BP 神经网络的合成算法，在训练数据中得到的头动参数允许范围内随机生成头动参数的方法（简称为随机头动方法）。

本文采用 HTS 工具包 http：// hts. sp. nitech. ac. jp 实现基于 Trajectory-HMM 的头动合成。对每个头动类别建立一个 5 流的 MSD（multi-space probability distribution）-HMM，其中：第 1 流为 12 维的 Smoothed MFCC 和 Energy 以及一阶、二阶差分共 39 维；第 2 流为 Pitch，第 3、4 流分别为 Pitch 的一阶和二阶差分，第 5 流为 9 维的头动特征。每个 HMM 模型设置为 7 个状态，每个状态的 Gauss 数目设置为经验值，模型训练采用训练集中 133 句音视频数据。BP 网络的输入节点为 42 维音频特征，输出为 3 维头动特征，中间隐层采用 80 个节点。网络的学习效率设置为 0. 001，训练次数为 3 000 次，训练算法为量化共轭梯度法。由于音频特征的频率是 100 Hz，头动特征的频率是 25 Hz，因此在建模前，将头动特征进行了重采样，使音视频具有相同的频率。

### 4.2 实验结果

头动类型的识别率对于 Trajectory-HMM 方法的合成效果至关重要，因此本文在 30 句测试句子上计算了识别率。手工标注 6 类、手工标注 4 类和自动聚 8 类方案的正确率分别为 50. 62%、55. 51%和 47. 09%。此结果是在调整模型设置下获得的最优结果。手工标注 4 类的正确率最高，自动聚 8 类的正确率最低。整体而言，识别率维持在较低水平，这与前期研究结果是一致的[9]。

为了衡量合成头动曲线和真实曲线之间的关系，本文采用了均方误差 MSE 和典型相关分析（canonical correlation analysis，CCA）作为客观评价评价标准。MSE 反映了预测误差，取值越小越好。CCA 是用来研究两组变量之间相关性的多元统计方法，其取值越大表明两组变量的相关性越大。在 30 句测试语句上计算的平均 MSE、CCA 结果总结于表 1。为了给出 Trajectory HMM 方法效果的上限，表 1 也给出了手工标注的头动类别信息（即识别率 100%）作为输入的结果。

表 1 各种头动合成方案的客观评测结果

| 方法 | 手工标注 4 类 | 手工标注 4 类（完全正确） | 手工标注 6 类 | 手工标注 6 类（完全正确） | 自动聚 8 类 | 自动聚 8 类（完全正确） | BP 神经网络 | 随机头动 |
|---|---|---|---|---|---|---|---|---|
| MSE | 0. 298 7 | 0. 233 5 | 0. 290 2 | 0. 233 9 | 0. 319 8 | 0. 304 7 | 0. 279 4 | 0. 372 9 |
| CCA | 0. 621 0 | 0. 640 3 | 0. 681 2 | 0. 684 9 | 0. 655 0 | 0. 868 4 | 0. 7471 | 0. 361 6 |

从表 1 中的实验结果可以看出，基于神经网络的头动合成方案产生的 MSE 最小，CCA 最大，说明该方案从语音中预测出的头动曲线和真实头动最为接近。随机头动方案的效果最差，预测误差最大，与真实头动的相关度不高。对比 3 种基于 Trajectory HMM 的方案，手工标注 4 类和手工标注 6 类的 MSE 相当，自动聚 8 类的 MSE 较大；而从 CCA 指标上看，手工标注 6 类的稍好。从表 1 中还可以看出，在头部姿态标注完全正确的情况下，3 种 Trajectory HMM 策略的 MSE 有所降低，但是仍然逊于神经网络方法；在 CCA 指标上，除自动聚 8 类的方案之外，其余两种方案都未能超过神经网络方法。可见，正是由于识别率不高，影响了 Trajectory HMM 方案的效果。图 4 展示了基于 BP 神经网络的头动合成方法在一个测试集句子上合成的点头（Pitch）曲线。可见，合成的点头曲线和实际点头趋势十分接近。

本文同时进行了主观评测。用 5 个方案合成的头动曲线，驱动一个 3-D 虚拟说话人的头动。采用文[17]中的方法配以同步的语音与唇动，不同方案生成的动画仅是头动不同，语音和唇动则保持一致。为了与真实头动进行对比，以 AAM-FPT 工具获得的头动参数作为第 6 种方案。将生成的头动，随机播放给 10 个受测试者，让受测试者根据头动的自然度进行打分（5 分制）。各方案的主观平均意见得分（mean opinion score，MOS）总结于表 2。可以看出，使用神经网络的合成效果，明显高于其他方案，略低于真实数据的效果；随机头动的效果最差。这与客观评测结果是一致的。同时也观察到，所有方案的整体 MOS 偏低，主要原因是采用的虚拟人仅有头部，未能体现头颈随身体的自然运动。这说明仅有自然的头动是不够的，躯干运动也是提高自然度的重要因素。

表 2 各种头动合成方案的主观平均意见得分 MOS

| 方法 | 手工标注 4 类 | 手工标注 6 类 | 自动聚 8 类 | 神经网络 | 随机头动 | 真实数据 |
|---|---|---|---|---|---|---|
| MOS | 2. 79 | 2. 96 | 2. 85 | 3. 2 | 2. 68 | 3. 65 |

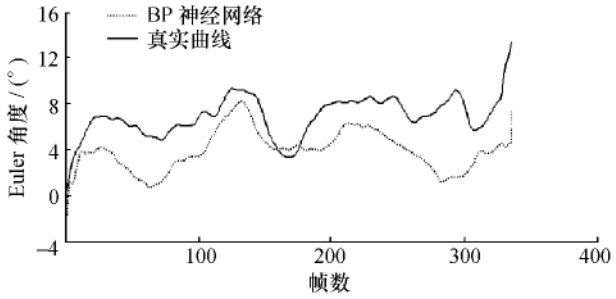


**图 4 BP 神经网络在一个句子上合成的点头(Pitch)曲线**

## 5 结 论

**本文研究了从语音信号预测伴随头动的方法。本文尝试了不同头动模式划分方法在基于 Trajectory HMM 头动合成上的效果。此类方法依赖于头动模式的正确识别。然而,语音和头动之间是非确定性的多对多映射关系,很难用类别描述清楚,因此识别率不高,头动合成效果受限。为此,本文尝试采用 BP 神经网络回归方法,通过学习语音与头动之间的映射关系,实现语音信号到头动参数之间的直接连续映射。实验表明,基于 BP 神经网络的回归方法明显优于基于 Trajectory HMM 方法,有效地提高了语音到头动预测的准确度和头动合成的自然度。**

## 参考文献 (References)

[1] **张申. 虚拟说话人可视表现力的研究 [D]. 北京: 清华大学**, 2011.
ZHANG Shen. Research on Visual Expressivity of Talking Avatar [D]. Beijing: Tsinghua University, 2011. (in Chinese)

[2] Feldman R S. Fundamentals of Nonverbal Behavior [M]. Cambridge, UK: Cambridge University Press, 1991.

[3] Munhall K, Jones J, Callan D, et al. Visual prosody and speech intelligibility: Head movement improves auditory speech perception [J]. *Psychological Science*, 2004, **15**(2): 133-137.

[4] McNeill D. Gesture and Thought [M]. Chicago, IL, USA, University of Chicago Press, 2005.

[5] Pelachaud C, Badler N, Steedman M. Generating facial expressions for speech [J]. *Cognitive Science: A Multidisciplinary Journal*, 1996, **20**(1): 1-46.

[6] ZHANG Shen, WU Zhiyong, Meng H, et al. Head movement synthesis based on semantic and prosodic features for a Chinese expressive avatar [C]// Proc ICASSP. Honolulu, HI, USA: IEEE Press, 2007: 837-840.

[7] Busso C, Deng Z, Grimm M, et al. Rigid head motion in expressive facial animation: Analysis and synthesis [J]. *IEEE Trans on Audio, Speech and Language Processing*, 2006, **15**(3): 1075-1086.

[8] Sargin M E, Erzin E, Yemez Y, et al. Prosody-driven head-gesture animation [C]// Proc ICASSP. Honolulu, HI, USA: IEEE Press, 2007: 677-680.

[9] Hofer G, Shimodaira H. Automatic head motion prediction from speech data [C]// Proc Interspeech. Grenoble, France: ISCA, 2007.

[10] Gregor H, Hiroshi S, Junichi Y. Speech driven head motion synthesis based on a trajectory model [C]// Proc Siggraph. San Diego, CA, USA: ACM Press, 2007.

[11] Binh L, Ma X, Deng Z. Live speech driven head-and-eye motion generators [J]. *IEEE Transactions on Visualization and Computer Graphics*, 2012, **18**(11): 1902-1914.

[12] Graf H P, Cosatto E, Strom V, et al. Visual prosody: Facial movements accompanying speech [C]// Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition. Orlando, FL, USA: IEEE Press, 2002: 381-386.

[13] Keiichi T, Takayoshi Y, Takashi M, et al. Speech parameter generation algorithms for HMM-based speech synthesis [C]// Proc ICASSP. Istanbul, Turkey: IEEE Press, 2000: 1315-1318.

[14] Hadar U, Steiner T, Grant E, et al. Head movement correlates of juncture and stress at sentence level [J]. *Language and Speech*, 1983, **2**: 451-471.

[15] Orozco J, Roca F X, Gonzlez J. Real-time gaze tracking with appearance-based models [J]. *Machine Vision and Applications*, 2008, **20**(6): 353-364.

[16] Hecht-Nielsen R. Theory of the back propagation neural network [C]// Proc IJCNN. Detroit, MI, USA: IEEE Press, 1989: 583-604.

[17] **李冰锋, 谢磊, 周祥增, 等. 实时语音驱动的虚拟说话人 [J]. 清华大学学报: 自然科学版**, 2011, **51**(9): 1180-1186.
LI Bingfeng, XIE Lei, ZHOU Xiangzeng. Real-time speech driven talking avatar [J]. *J Tsinghua Univ: Sci and Tech*, 2011, **51** (9): 1180-1186. (in Chinese)