融合人脸表情的手语到汉藏双语情感语音转换

宋 南，吴沛文，杨鸿武
(西北师范大学物理与电子工程学院，甘肃兰州 730070)

摘要：针对聋哑人与正常人之间存在的交流障碍问题，提出了一种融合人脸表情的手语到汉藏双语情感语音转换的方法。首先使用深度置信网络模型得到手势图像的特征信息，并通过深度神经网络模型得到人脸信息的表情特征。其次采用支持向量机对手势特征和人脸表情特征分别进行相应模型的训练及分类，根据识别出的手势信息和人脸表情信息分别获得手势文本及相应的情感标签。同时，利用普通话情感训练语料，采用说话人自适应训练方法，实现了一个基于隐 Markov 模型的情感语音合成系统。最后，利用识别获得的手势文本和情感标签，将手势及人脸表情转换为普通话或藏语的情感语音。客观评测表明，静态手势的识别率为 92.8%，在扩充的 Cohn-Kanade 数据库和日本女性面部表情(Japanese Female Facial Expression, JAFFE)数据库上的人脸表情识别率为 94.6%及 80.3%。主观评测表明，转换获得的情感语音平均情感主观评定得分 4.0 分，利用三维情绪模型(Pleasure-Arousal-Dominance, PAD)分别评测人脸表情和合成的情感语音的 PAD 值，两者具有很高的相似度，表明合成的情感语音能够表达人脸表情的情感。

关键词：手势识别；表情识别；深度神经网络；汉藏双语情感语音合成；手语到语音转换
中图分类号：TP391 **文献标识码：**A **文章编号：**1000-3630(2018)-04-0372-08
DOI 编码：10.16300/j.cnki.1000-3630.2018.04.014

Gesture-to-emotional speech conversion based on gesture recognigion and facial expression recognition

SONG Nan, WU Pei-wen, YANG Hong-wu
(*College of Physics and Electronic Engineering*, *Northwest Normal University*, *Lanzhou* 730070, *Gansu*, *China*)

Abstract: This paper proposes a face expression integrated gesture-to-emotional speech conversion method to solve the communication problems between healthy people and speech disorders. Firstly, the feature information of gesture image are obtained by using the model of the deep belief network (DBN) and the features of facial expression are extracted by a deep neural network (DNN) model. Secondly, a set of support vector machines (SVM) are trained to classify the gesture and facial expression for recognizing the text of gestures and emotional tags of facial expression. At the same time, a hidden Markov model-based Mandarin-Tibetan bilingual emotional speech synthesis is trained by speaker adaptive training with a Mandarin emotional speech corpus. Finally, the Mandarin or Tibetan emotional speech is synthesized from the recognized text of gestures and emotional tags. The objective tests show that the recognition rate for static gestures is 92.8%. The recognition rate of facial expression achieves 94.6% on the extended Cohn-Kanade database (CK+) and 80.3% on the JAFFE database respectively. Subjective evaluation demonstrates that synthesized emotional speech can get 4.0 of the emotional mean opinion score. The pleasure-arousal-dominance (PAD) tree dimensional emotion model is employed to evaluate the PAD values for both facial expression and synthesized emotional speech. Results show that the PAD values of facial expression are close to the PAD values of synthesized emotional speech. This means that the synthesized emotional speech can express the emotion of facial expression.

Key words: gesture recognition; facial expression recognition; deep neural network; Mandarin-Tibetan bilingual emotional speech synthesis; gesture to speech conversion

0 引 言

手语是目前言语障碍者与正常人之间最重要的一种沟通方式，手语识别研究一直受到广泛的关注[1]，手势识别技术逐渐成为人机交互系统方面的研究热点。早期，利用穿戴技术通过数据手套进行手语识别[2]。近年来，模式识别技术中的隐 Markov 模型(Hidden Markov Model, HMM)[3]、反向传播(Back Propagation, BP)神经网络[4]及支持向量机(Support Vector Machine, SVM)[5]等算法应用在手势识别上，获得了一定的效果。目前，随着深度学习技术的发展，深度学习也应用到手语识别中[6]，使

收稿日期：2017-10-09；**修回日期：**2017-12-17
基金项目：国家自然科学基金(11664036、61263036、61262055)、甘肃省高等学校科技创新团队项目(2017C-03)资助。
作者简介：宋南(1990－)，男，河北迁安人，硕士研究生，研究方向为信号与信息处理。
通讯作者：杨鸿武，E-mail: yanghw@nwnu.edu.cn

得手语识别率获得了较大提高。同时，在日常生活交往中，面部表情在言语障碍者的交流中也起到很重要的作用，表情可以让交流的信息传达得更加准确。现有的表情识别技术发展迅速，基于 SVM[7]、Adaboost[8]、局部二值模式(Local Binary Pattern, LBP)、主成分分析(Principal Components Analysis, PCA)[9]以及深度学习的人脸表情识别[10]都已经得到了实现。手语信息与人脸表情信息的融合将会让信息表达更加明确。目前基于 HMM 的语音合成方法广泛应用在情感语音合成领域[11-12]，通过该方法可将文本信息转换成情感语音。但现有的研究方向大都是分别对手势、人脸表情及情感语音合成进行研究。一些学者采用信息融合的方法，将人脸表情、肢体语言及语音信息进行融合，实现了多模式融合下的情感识别[13]；将手势识别与语音信息融合，实现了对机器人的指挥[14]；将面部表情信息与语音信息融合，实现了对机器人轮椅导航的控制[15]；这些研究表明，多模式信息融合逐渐成为一种趋势。前期的研究[16-17]虽然实现了手语到语音的转换，但合成出的语音并没有包含感情和情绪的变化，忽视了聋哑人情感的语音表达，容易使听者的理解产生歧义。

将手语和人脸表情的识别技术与情感语音合成方法相结合，实现融合人脸表情的手语到情感语音的转换，对言语障碍者的日常交流具有重要作用。本文首先利用静态手势识别获得手势表达的文本，利用人脸表情识别获得表达的情感信息。同时以声韵母作为语音合成基元，实现了一个基于 HMM 的汉藏双语情感语音合成，将识别获得的手势文本和情感信息转换为相应的普通话或藏语情感语音。

1　系统框架

融合人脸表情的手语到汉藏双语情感语音转换系统框架如图 1 所示。为了实现转换系统，将系统设计为三部分：手势和人脸表情的识别、情感语音声学模型训练及情感语音合成。在识别阶段，将输入的手势图像进行预处理，再通过深度置信网络(Deep Belief Network, DBN)模型进行特征提取得到手势特征，利用 SVM 识别得到手势种类；将输入的人脸表情图像进行预处理，再通过深度神经网络(Deep Neural Network, DNN)模型进行特征提取得到表情特征，利用 SVM 识别得到情感标签。在训练阶段，将语料库中的语音和文本分别进行参数提取与文本分析，得到声学参数和标注信息，再通过情感语音的合成平台进行 HMM 训练，得到不同情感的语音声学模型。在合成阶段，将获得的手势种类利用定义好的手势文本字典得到手势文本，通过文本分析得到情感语音合成所需的上下文相关的标注信息，同时利用情感标签选择情感语音声学模型，最终将上下文相关的标注信息和情感语音声学模型，通过情感语音合成系统合成出情感语音。

2　融合人脸表情的手语到情感语音合成

2.1　手势识别

手势识别主要包括 3 个部分：预处理、特征提取以及 SVM 识别。图像的预处理过程通过对手势信息进行数据整合，把采集到的手势图像转化为灰

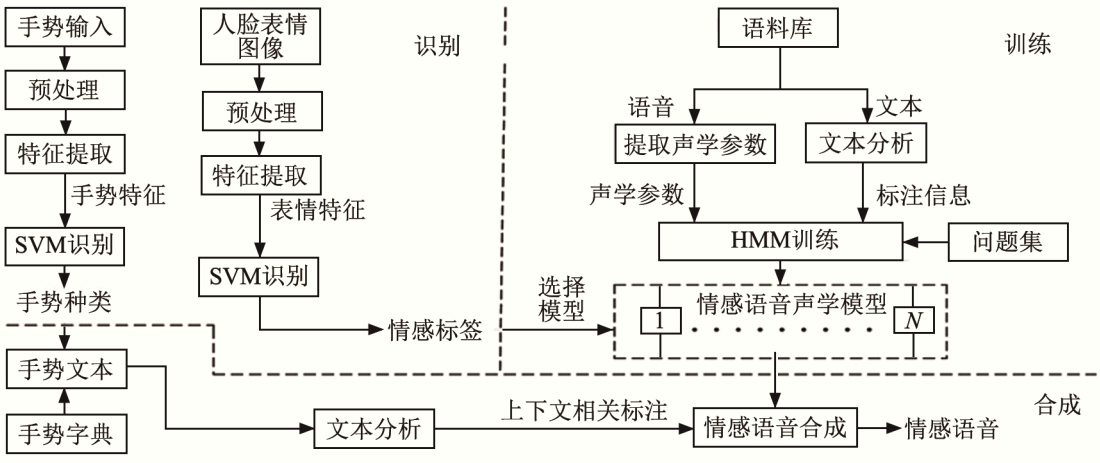

图 1　面向言语障碍者的手语到情感语音转换系统框架

Fig.1　Block diagram of gesture-to-emotional speech conversion for speech disorder

度图像，并将其格式从 28×28 变换为 784×1。针对所有图像构成一个二维矩阵，然后构建数据立方体。(x 轴坐标表示一个小组内不同样本的编号，y 轴坐标表示一个小组中特定一个样本的维度，z 轴表示小组的个数)，把其作为 DBN 模型统一读入数据的格式。手势特征采用 5 层的 DBN 模型进行提取，其过程包括受限玻尔兹曼机(Restricted Boltzmann Machine, RBM)调节和反馈微调，利用 RBM 来调节相邻两层之间的权值[18]，RBM 在隐藏层到可见层之间有连接，每层内部都没有连接，其隐藏层与可见层之间的关系可以用能量函数表示为

$$E(v,h;\theta)=-\sum_{i=1}^{V}\sum_{j=1}^{H}w_{ij}v_ih_j-\sum_{i=1}^{V}b_iv_i-\sum_{j=1}^{H}a_jh_j \quad (1)$$

其中，v_i 和 h_i 分别表示可见层节点与隐藏层节点的状态，取值一般为 0 或 1，$\theta=\{W,a,b\}$ 是模型参数；b_i 和 a_j 分别代表可见层神经元和隐藏层神经元相应的偏置，w_{ij} 表示连接权重。可见层与隐藏层的联合分布为

$$p(v,h)=\mathrm{e}^{-E(v,h)}/\sum_{v,h}\mathrm{e}^{-E(v,h)} \quad (2)$$

可见层与隐藏层之间的条件概率计算如下：

$$p(h_j=1|v)=\mathrm{sigm}(\sum_{i=1}^{V}w_{ij}v_i+a_j) \quad (3)$$

$$p(v_j=1|h)=\mathrm{sigm}(\sum_{j=1}^{H}w_{ij}h_j+b_i) \quad (4)$$

其中，$\mathrm{sigm}(z)=(1+\mathrm{e}^{-z})^{-1}$ 是 sigmoid 函数，是一种神经元非线性函数。RBM 模型的更新权重能够通过导数概率的对数得到。

在调节过程中，通过逐层训练的方式得到每层的权值，完成可见层与隐藏层之间的反复三次转换，分别得到相应的重构目标，并利用缩小原对象同重构对象之间的差异，实现对 RBM 参数的调节。

微调是把全部的经过初始化后的 RBM 按训练的顺序串联起来，组成一个深度置信网络，通过深度模型的反馈微调可以得到手势图像的特征信息。SVM 识别过程是把获得的手势图像的特征信息进行分类识别得到手势种类，其过程如图 2 所示。

2.2 人脸表情识别

人脸表情识别过程如图 3 所示，包括预处理、特征提取和 SVM 识别 3 个阶段。预处理阶段对原始图像中可能会影响到特征提取结果的一些不重要的背景信息进行处理。首先对原始的输入图像使用具有 68 个面部地标点的检测器进行检测，然后再将图像调整到地标边缘，在保留完整表情信息的前提下对图像进行裁剪，剪裁后删除图像的一些没有特定信息的部分，使神经网络模型的输入图像大小为 96×96。在特征提取阶段，利用一个 22 层的 DNN 模型进行特征提取，从输入的每张表情图像中得到 128 维的特征。在 SVM 识别阶段，将得到的表情特征利用一个训练好的 SVM 分类器进行分类识别，从而得到人脸表情对应的情感标签。

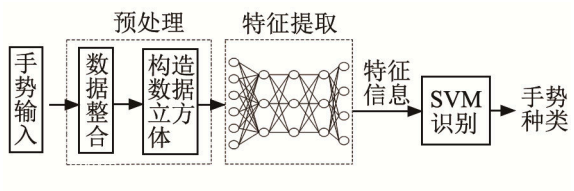

图 2 手势识别
Fig.2 Gesture recognition

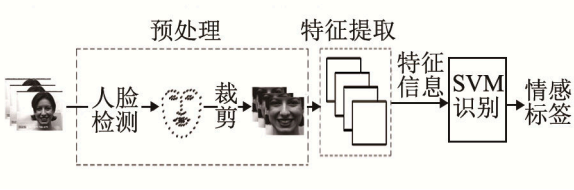

图 3 人脸表情识别
Fig.3 Facial expression recognition

2.3 情感语音声学模型训练

本文以普通话和藏语的声母和韵母为语音合成的基本单元，利用说话人自适应训练(Speaker Adaptive Training, SAT)获得了情感的语音声学模型，情感语音声学模型的训练过程如图 4 所示。

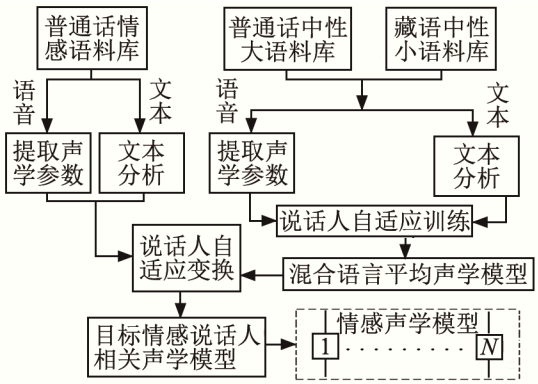

图 4 情感语音声学模型训练过程
Fig.4 Training process of emotional speech acoustic mode

首先，利用一个普通话中性大语料库(多说话人)和一个藏语中性小语料库(1 个说话人)中的语音和文本分别进行声学参数提取与文本分析，得到声学特征对数基频(Log-fundamental Frequency, logF0)和广义梅尔倒谱系数(Mel-generalized Cepstral, MGC)以及文本的标注信息(上下文相关标注和单音素标注)；然后利用声学特征和标注信息进行说话人自适应训练，得到混合语言平均声学模型。

最后将从多说话人普通话情感语料库中提取

的情感语音声学特征和相应文本的标注信息，与获得的平均声学模型一起通过说话人自适应变换得到目标情感的说话人相关声学模型，以合成普通话或藏语的情感语音。

本文采用基于半隐马尔可夫模型(Hidden Semi-Markov Model, HSMM)[19]的说话人自适应训练算法训练声学模型，以减少不同说话人之间的差异对合成语音音质的影响。时长分布与状态输出分布的线性回归方程分别为

$$\hat{o}_i^s(t)=A^s o_i(t)+b^s=W_i^s(t)\xi_i(t) \tag{5}$$

$$\hat{d}_j^s(t)=\alpha^s d_j(t)+\beta^s=X_j^s(t)\xi_j(t) \tag{6}$$

其中，$\hat{o}_i^s(t)$ 与 $\hat{d}_j^s(t)$ 分别为训练说话人 s 的状态输出均值向量和状态时长均值向量，$\boldsymbol{\xi}_i=[o_i,1]^{\mathrm{T}}$。$\boldsymbol{W}=[A,b]$ 为训练说话人 s 和平均声学模型之间的状态输出分布，$\boldsymbol{X}=[\alpha,\beta]$ 为其状态时长分布差异的变换矩阵，$\boldsymbol{\xi}_j=[d_j,1]^{\mathrm{T}}$。$o_i$ 与 d_j 分别代表平均观测向量和平均时长。

本文采用约束最大似然线性回归(Constrained Maximum Likelihood Linear Regression, CMMLR)[20]训练得到平均声学模型，进而获得上下文相关的多空间分布半隐马尔科夫模型(Multi-Space Hidden semi-Markov models, MSD-HSMM)。训练平均声学模型后，将基于 MSD-HSMM 的 CMMLR 自适应算法应用于多说话人普通话情感语料库，得到用来合成普通话情感语音和藏语情感语音的说话人相关混合语言目标情感声学模型。状态 i 下状态时长 d 和特征向量 o 的变换方程如式(7)、(8)所示：

$$p_i(d)=N\left(d;\alpha m_i-\beta,\alpha\sigma_i^2\alpha\right)=\left|\alpha^{-1}\right|N\left(\alpha\boldsymbol{\psi};m_i,\sigma_i^2\right) \tag{7}$$

$$b_i(o)=N\left(o;\boldsymbol{A}u_i-\boldsymbol{b},\boldsymbol{A}\sum\nolimits_i\boldsymbol{A}^{\mathrm{T}}\right)=\left|\boldsymbol{A}^{-1}\right|N\left(\boldsymbol{W}\boldsymbol{\xi};u_i,\sum\nolimits_i\right) \tag{8}$$

其中，$\boldsymbol{\psi}=[d,1]^{\mathrm{T}}$，$\boldsymbol{\xi}=[o^{\mathrm{T}},1]$，$m_i$ 为时长分布均值，u_i 为状态输出分布均值，σ_i^2 为方差，$\boldsymbol{\Sigma}_i$ 为对角协方差矩阵，$\boldsymbol{X}=[\alpha^{-1},\beta^{-1}]$ 为状态时长概率密度分布的变换矩阵，$\boldsymbol{W}=[\boldsymbol{A}^{-1},\boldsymbol{b}^{-1}]$ 为目标说话人状态输出概率密度分布的线性变换矩阵。

通过基于 HSMM 的自适应变换算法，可对语音数据的基频、频谱和时长参数进行变换和归一化。对于长度为 T 的自适应数据 O，可通过变换 $\boldsymbol{\Lambda}=(\boldsymbol{W},\boldsymbol{X})$ 进行最大似然估计：

$$\tilde{\boldsymbol{\Lambda}}=\left(\tilde{\boldsymbol{W}},\tilde{\boldsymbol{X}}\right)=\arg\max_{\boldsymbol{\Lambda}}P(O|\lambda,\boldsymbol{\Lambda}) \tag{9}$$

其中，λ 为 HSMM 的参数集。

最后采用最大后验概率(Maximum A Posteriori, MAP)算法[21]对目标语言目标情感的特定人模型进行更新与修正。对于给定的模型 λ，若其前向概率和后向概率分别为：α_t^i 和 β_t^i，则其在状态 i 下连续观测序列 $o_{t-d+1}\cdots o_t$ 的概率 $k_t^d(i)$ 为

$$k_t^d(i)=\frac{1}{P(O|\lambda)}\sum_{\substack{j=1\\ j\neq i}}^{N}\alpha_{t-d}(j)p(d)\prod_{s=t-d+1}^{t}b_i(o_s)\beta_t(i) \tag{10}$$

MAP 估计为

$$\hat{m}_i=\frac{\tau\bar{m}_i+\sum_{t=1}^{T}\sum_{d=1}^{t}K_t^d(i)d}{\tau+\sum_{t=1}^{T}\sum_{d=1}^{t}K_t^d(i)d} \tag{11}$$

$$\hat{u}_i=\frac{\omega\bar{u}_i+\sum_{t=1}^{T}\sum_{d=1}^{t}K_t^d(i)\sum_{s=t-d+1}^{t}o_s}{\omega+\sum_{t=1}^{T}\sum_{d=1}^{t}K_t^d(i)d} \tag{12}$$

其中，$\bar{m}_i$ 和 $\bar{u}_i$ 为线性回归变换后的均值向量，τ 和 ω 分别为时长分布和状态输出的 MAP 估计参数。$\hat{m}_i$ 和 $\hat{u}_i$ 分别为自适应向量 $\bar{m}_i$ 和 $\bar{u}_i$ 的加权平均 MAP 估计值。

2.4　手语到情感语音转换

为了获得手势文本，根据《中国手语》[22]中定义的手势种类的含义，设计了一个手势字典，该字典给出了每个手势对应的语义文本。在手语到情感语音的转换过程中，首先通过手势识别获得手势类别，然后查找手势字典，获得手势文本，最后对手势文本进行文本分析，获得文本的声韵母信息以及声韵母的上下文信息，从而能够利用决策树选择出最优的声韵母的声学模型。声韵母的上下文信息以上下文相关标注的形式给出，包括普通话或藏语的声韵母信息、音节信息、词信息[23]、韵律词信息[24]、短语信息和语句信息。同时，采用人脸表情识别获得情感标签，利用情感标签选择相应情感的语音声学模型，从而能够利用文本的上下文相关标注信息合成出普通话或藏语的情感语音。手语到情感语音转换流程如图 5 所示。

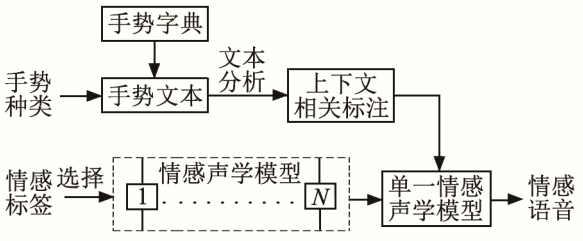

图 5　手语到情感语音转换框图

Fig.5　Block diagram of sign language to emotional speech conversion

3　实验结果

3.1　手势识别

3.1.1　手势数据

在实验中构造的手势样本集合主要来自 2 位测试人所生成的样本，每位测试人打 30 种手势，每种手势的样本个数均为 1 000，以此来生成 30 个深

度学习模型。预定义的 30 种静态手势如图 6 所示。

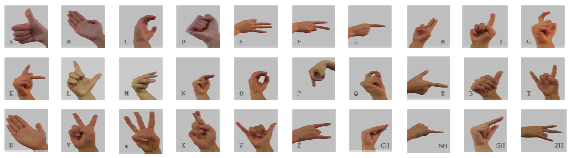

图 6 预定义的 30 种手势
Fig.6 Thirty kinds of static sign languages

3.1.2 手势识别率

为了验证 DBN 模型在手势识别上的有效性，本文从图 6 所示的 30 种手势库中随机挑选了 4 000 个样本，分别利用 DBN 模型和 PCA 方法进行了 5 次交叉实验，每次实验的训练集和测试集样本数分别为 3 200 和 800，并将这五次实验分别进行编号(集 1 到集 5)；最终利用 SVM 识别得到如表 1 所示的识别率。从表 1 中可以看出，在 5 次交叉验证中，利用 DBN 模型进行特征提取的手势识别率优于 PCA 方法，表明通过 DBN 模型提取到的特征能更好地反映出手势的本质特征。

表 1 5 次交叉验证识别率(%)
Table 1 Recognition rates of five cross validation(%)

类别	集 1	集 2	集 3	集 4	集 5
DBN	91.3	94.7	92.5	92.2	93.1
PCA	86.7	89.1	87.5	86.9	88.4

3.2 人脸表情识别

3.2.1 人脸表情库数据

本文采用扩充的 Cohn-Kanade 数据库(the extended Cohn-Kanade database, CK+)[25]和日本女性面部表情(Japanese Female Facial Expression, JAFFE)数据库[26]进行人脸表情的训练和测试。CK+数据库中每个序列图像都是以中性表达式开始到情感峰值结束。实验数据库中包含 8 种情感类别的表情图像，但在实验中，蔑视和中性表情图像没有被使用，并且只选取了一些具有明显表情特征信息的图像来作为样本集使用。将 JAFFE 数据库中 7 种表情中的 6 种表情进行了实验，没有使用中性表情图像，其中每人的一种表情图像大小均为 256×256。数据库中图像的一些例子如图 7 所示。

3.2.2 DNN 模型

本文采用了 nn4.small2 的神经网络模型[27]去提取表情图像特征，图 8 展示了一张裁剪后的图像经过该模型的第一层卷积后输出的特征图，该图显示了输入图像的第一个卷积层的 64 个全部滤镜。网络模型定义如表 2 所示。其中包含了 8 个 Inception 的模块。池化层可以有效地缩小矩阵的尺寸，而最大池化表示对邻域内特征点取最大，平均池化表示对邻域内特征点只求平均。池项目表示嵌入的最大

图 7 数据库示例
Fig.7 Database example

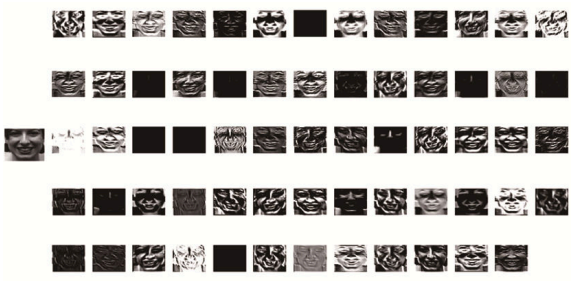

图 8 卷积层可视化示例
Fig.8 An example of convolution layer visualization

表 2 网络模型定义
Table 2 Definition of deep neural network

层	输出尺寸	#1×1	#3×3 缩减
卷积 1 (7×7×3,2)	48×48×64	/	/
最大池化+标准	24×24×64	/	/
inception(2)	24×24×192	/	64
标准+最大池化	12×12×192	/	/
inception (3a)	12×12×256	64	96
inception (3b)	12×12×320	64	96
inception (3c)	6×6×640	/	128
inception (4a)	6×6×640	256	96
inception (4e)	3×3×1 024	/	160
inception (5a)	3×3×736	256	96
inception (5b)	3×3×736	256	96
平均池化	736	/	/
线性	128	/	/
ℓ_2 标准化	128	/	/

层	#3×3	#5×5 缩减	#5×5	池项目
卷积 1 (7×7×3,2)	/	/	/	/
最大池化+标准	/	/	/	*m* 3×3, 2
inception(2)	192	/	/	/
标准+最大池化	/	/	/	*m* 3×3, 2
inception (3a)	128	16	32	*m*, 32*p*
inception (3b)	128	32	64	ℓ_2, 64*p*
inception (3c)	256,2	32	64,2	*m* 3×3, 2
inception (4a)	192	32	64	ℓ_2, 128*p*
inception (4e)	256,2	64	128,2	*m* 3×3, 2
inception (5a)	384	/	/	ℓ_2, 96*p*
inception (5b)	384	/	/	*m*, 96*p*
平均池化	/	/	/	/
线性	/	/	/	/
ℓ_2 标准化	/	/	/	/

池化之后的投影层中 1×1 过滤器的个数，池项目中最大池化用 m 表示，降维后的池化用 p 表示。

3.2.3　表情识别率

在 CK+数据库上进行 5 次交叉验证的实验，得到 6 种表情相应的识别率。在 JAFFE 数据库上进行 3 次交叉验证的实验，得到 6 种表情相应的识别率。如表 3 所示。

从表 3 可以看出，JAFFE 的数据库上的识别率要低于 CK+数据库上的识别率，主要原因是在实验中 JAFFE 数据库的表情图片数量少于 CK+数据库的表情图片数量。

表 3　不同数据库上的人脸表情识别率(%)

Table 3　Facial expression recognition rates under differenet databases (%)

类别	CK+	JAFFE
愤怒	96.7	86.6
厌恶	93.1	82.7
恐惧	97.6	75.0
快乐	98.3	80.6
悲伤	92.9	83.8
惊讶	89.0	73.3
平均识别率	94.6	80.3

3.3　情感语音合成

3.3.1　语　料

普通话语料库选用 7 个女性说话人的中性语料，每个说话人的语料各包含 169 句，共计 1 183 句(7×169 句)语料。普通话情感语料库，是本研究设置特定的场景采用激发引导方式录制的 9 个女性说话人 11 种情感的普通话情感语音库，每个说话人的每种情感语料各包含 100 句，录音人不是专业演员，实验中选取了其中的 6 种情感语料(9 人×6 种情感×100 句)。藏语语料库是本研究录制的一个藏语女性说话人的 800 句语料。所有实验的语音均采用 16 bit 量化、16 kHz 采样、单通道的 WAV 文件格式。采用 5 状态的上下文相关的一阶 MSD-HSMM 模型来建立声学模型。

3.3.2　情感相似度评测

通过情感平均意见得分(Emotional Mean Opinion Score, EMOS)，对合成的普通话情感语音以及藏语情感语音分别进行情感相似度评测。给 10 名普通话评测者播放 100 句原始普通话情感语音作为参考，然后按照情感顺序依次播放 6 种情感的普通话情感语音。同时给 10 名藏语评测者播放 100 句合成的中性藏语语音，作为中性参考语音，之后按照 6 种情感顺序播放藏语情感语音。在评测打分过程中是按照播放语音的先后顺序来进行的，要求评测者参照现实生活中的情感表达经验，给每句合成出的语音，按 5 分制进行情感相似度打分，结果如图 9 所示。

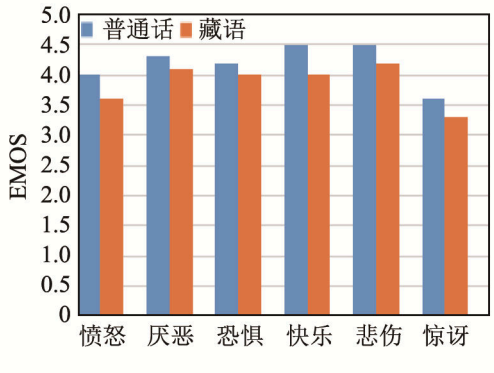

图 9　合成普通话和藏语的情感语音 EMOS 得分

Fig.9　The EMOS scores of synthesized emotional Tibetan and Mandarin speeches

从图 9 中可以看出，利用普通话情感语料训练的情感声学模型合成出的藏语情感语音的 EMOS 评分，要低于合成出的普通话情感语音的 EMOS 评分。

3.3.3　客观评测

由于只有普通话情感语料库，所以仅对合成的普通话情感语音进行了客观评测。本文计算了原始语音与合成语音在时长、基频及谱质心上的均方根误差(Root Mean Square Error, RMSE)，结果如表 4 所示。从表 4 可以看出，时长、基频及谱质心的均方根误差值较小，说明合成的普通话情感语音与原始的普通话情感语音比较接近，合成的情感语音音质较好。

表 4　普通话合成情感语音与原始情感语音在时长、基频及谱质心上的均方根误差

Table 4　RMSEs of duration, fundamental frequency and spectral centroid between the synthesized and the original emotional Mandarin speeches

类型	时长 RMSE/s	基频 RMSE/Hz	谱质心 RMSE/Hz
愤怒	0.59	53.59	376.32
厌恶	0.42	46.40	490.88
恐惧	0.52	35.55	359.23
快乐	0.39	46.71	387.91
悲伤	0.72	16.79	306.32
惊讶	0.32	42.16	391.75
平均	0.49	40.20	385.40

3.4　表情图片与情感语音的 PAD 评测

为了进一步评测合成语音对原始人脸表情的情感表达程度，本文采用 PAD 三维情绪模型，对比了表情图片的PAD 值与合成语音的PAD 值的差异。

本文采用简化版本的 PAD 情感量化表[28]，对人脸表情图片及其对应的情感语音在PAD的3个情绪维度上进行评分。首先随机播放所有人脸表情图片，评测者根据观测到图片时感受到的心理情绪状态，完成 PAD 情绪量表。然后随机播放合成的情感语音，同样要求评测者根据听情感语音时感受到的心理情绪状态，完成 PAD 情绪量表。由于藏语评测人不足，所以本文只对合成的普通话情感语音进行了 PAD 评测。最后，计算出在同一种情感状态下表情图片的 PAD 值与情感语音的 PAD 值的欧氏距离。评测结果如表 5 所示。从表 5 可以看出，表情图片和情感语音的 PAD 值在同一情感状态下的欧氏距离较小，表明合成的情感语音能够较为准确地再现人脸表情的情感状态。

表 5 PAD 的评测结果

Table 5 PAD evaluation for facial expression and synthesized emotional speech

情感类型	表情图片评价结果([-1,1])			
	P	A	D	
愤怒	–0.84	0.64	0.82	
厌恶	–0.50	0.26	0.39	
恐惧	–0.39	0.62	–0.63	
喜悦	0.67	0.78	0.37	
悲伤	–0.24	–0.40	–0.75	
惊讶	0.23	0.60	0.03	
情感类型	情感语音评价结果([–1,1])			欧氏距离
	P	A	D	
愤怒	–0.83	0.60	0.84	0.05
厌恶	–0.46	0.19	0.39	0.08
恐惧	–0.35	0.67	–0.67	0.08
喜悦	0.68	0.83	0.50	0.14
悲伤	–0.26	–0.35	–0.72	0.06
惊讶	0.26	0.70	0.03	0.10

4 结 论

本文提出了一种融合人脸表情的手语到汉藏双语情感语音转换的实现方法。首先，将手势库中的手势图像通过 DBN 模型进行特征提取，同时对人脸表情数据库(CK+和 JAFFE)中的表情图像利用 DNN 模型进行特征提取，把获得的手势特征与表情特征进行 SVM 识别，并分别转换为手势文本的上下文相关标注及相应的情感标签。再利用情感语料库以及中性语料库(普通话中性大语料库和藏语中性小语料库)，训练了一个基于 HMM 的普通话/藏语的情感语音合成器。最后，根据识别获得的情感标签选择的情感语音声学模型和手势文本的上下文相关标注进行情感语音合成，从而实现手势到情感语音的转换。实验结果表明，转换获得的汉藏双语情感语音的平均 EMOS 得分为 4.0 分；同时，利用 PAD 三维情绪模型对表情图片以及合成出的情感语音进行 PAD 评定后发现，表情图片与合成出的情感语音在 PAD 值上的欧式距离较小，表明合成的情感语音能够表达人脸表情的情感状态。进一步的工作将结合深度学习优化手势识别、人脸表情识别及汉藏双语情感语音合成的算法结构，提高识别率和合成情感语音的音质。

参 考 文 献

[1] KALSH E A, GAREWAL N S. Sign language recognition system[J]. International Journal of Computational Engineering Research, 2013, **3**(6): 15-21.

[2] ASSALEH K, SHANABLEH T, ZOUROB M. Low complexity classification system for glove-based arabic sign language recognition[C]//Neural Information Processing. Springer Berlin/Heidelberg, 2012: 262-268.

[3] GODOY V, BRITTO A S, KOERICH A, et al. An HMM-based gesture recognition method trained on few samples[C]// 2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2014: 640-646.

[4] YANG Z Q, SUN G. Gesture recognition based on quantum-behaved particle swarm optimization of back propagation neural network[J]. Computer application, 2014, **34**(S1): 137-140.

[5] GHOSH D K, ARI S. Static Hand Gesture Recognition using Mixture of Features and SVM Classifier[C]// 2015 Fifth International Conference on Communication Systems and Network Technologies (CSNT). IEEE, 2015: 1094-1099.

[6] OYEDOTUN O K, KHASHMAN A. Deep learning in vision-based static hand gesture recognition[J]. Neural Computing and Applications, 2017, **28**(12): 3941-3951.

[7] HSIEH C C, HSIH M H, JIANG M K, et al. Effective semantic features for facial expressions recognition using svm[J]. Multimedia Tools and Applications, 2016, **75**(11): 6663-6682.

[8] PRABHAKAR S, SHARMA J, GUPTA S. Facial Expression Recognition in Video using Adaboost and SVM[J]. Polish Journal of Natural Sciences, 2014, **3613**(1): 672-675.

[9] ABDULRAHMAN M, GWADABE T R, ABDU F J, et al. Gabor wavelet transform based facial expression recognition using PCA and LBP[C]//Signal Processing and Communications Applications Conference (SIU), 2014 22nd. IEEE, 2014: 2265-2268.

[10] ZHAO X, SHI X, ZHANG S. Facial expression recognition via deep learning[J]. IETE Technical Review, 2015, **32**(5): 347-355.

[11] BARRA-CHICOTE R, YAMAGISHI J, KING S, et al. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech[J]. Speech Communication, 2010, **52**(5): 394-404.

[12] WU P, YANG H, GAN Z. Towards realizing mandarin-tibetan bi-lingual emotional speech synthesis with mandarin emotional training corpus[C]//International Conference of Pioneering Computer Scientists, Engineers and Educators. Springer, Singapore, 2017: 126-137.

[13] CARIDAKIS G, CASTELLANO G, KESSOUS L, et al. Multimodal emotion recognition from expressive faces, body gestures and speech[C]// IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, Boston, MA, 2007: 375-388

[14] BURGER B, FERRANÉ I, LERASLE F, et al. Two-handed gesture recognition and fusion with speech to command a robot[J].

Autonomous Robots, 2012, **32**(2): 129-147.

[15] SINYUKOV D A, LI R, OTERO N W, et al. Augmenting a voice and facial expression control of a robotic wheelchair with assistive navigation[C]// 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE, 2014: 1088-1094.

[16] YANG H, AN X, PEI D, et al. Towards realizing gesture-to-speech conversion with a HMM-based bilingual speech synthesis system[C]// 2014 IEEE International Conference on Orange Technologies (ICOT). IEEE, 2014: 97-100.

[17] AN X, YANG H, GAN Z. Towards realizing sign language-to-speech conversion by combining deep learning and statistical parametric speech synthesis[C]// International Conference of Young Computer Scientists, Engineers and Educators. Springer Singapore, 2016:678-690.

[18] FENG F, LI R, WANG X. Deep correspondence restricted Boltzmann machine for cross-modal retrieval[J]. Neurocomputing, 2015, **154**: 50-60.

[19] ZEN H, TOKUDA K, BLACK A W. Statistical parametric speech synthesis[J]. Speech Communication, 2009, **51**(11): 1039-1064.

[20] YAMAGISHI J, KOBAYASHI T, NAKANO Y, et al. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2009, **17**(1): 66-83.

[21] SIOHAN O, MYRVOLL T A, LEE C H. Structural maximum a posteriori linear regression for fast HMM adaptation[J]. Computer Speech & Language, 2002, **16**(1): 5-24.

[22] 中国聋人协会. 中国手语[M]. 北京：华夏出版社, 2003.
China Association of the Deaf and Hard of Hearing. Chinese Sign Language[M]. Beijing: Huaxia Publishing House, 2003.

[23] YANG H, OURA K, WANG H, et al. Using speaker adaptive training to realize Mandarin-Tibetan cross-lingual speech synthesis[J]. Multimedia Tools & Applications, 2015, **74**(22): 9927-9942.

[24] 杨鸿武, 朱玲. 基于句法特征的汉语韵律边界预测[J]. 西北师范大学学报(自然科学版), 2013, **49**(1): 41-45.
YANG Hongwu, ZHU Ling. Predicting Chinese prosodic boundary based on syntactic features[J]. Journal of Northwest Normal University (Natural Science Edition), 2013, **49**(1): 41-45.

[25] LUCEY P, COHN J F, KANADE T, et al. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression[C]// Computer Vision and Pattern Recognition Workshops. IEEE, 2010:94-101.

[26] LYONS M, AKAMATSU S, KAMACHI M, et al. Coding facial expressions with gabor wavelets[C]// Third IEEE International Conference on Automatic Face and Gesture Recognition. IEEE, 1998: 200-205.

[27] AMOS B, LUDWICZUK B, SATYANARAYANAN M. OpenFace: A general-purpose face recognition library with mobile applications[R]. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[28] LI X M, FU X L, DENG G F. Preliminary application of the abbreviated PAD emotion scale to Chinese undergraduates[J]. Chinese Mental Health Journal, 2008, **22**(5): 327-329.