汉语音节与口形关系的研究

王志明　蔡莲红

清华大学计算机科学语技术系(100084)

摘要：对于说话者人脸合成和视频音频综合编码，由语音或文本预测说话者口形是很重要的。通过对汉语发音及其相关图像的研究，本文提出一种汉语声母韵母发音进行分类的方法，按类建立了与口形的对应关系，并得出各类声母的基本口形和韵母的口形变化曲线；另外，我们采用多层前馈神经网络实现了由语音信号倒谱系数及能量到部分口形参数的转换。

关键词：视觉语音，多层感知器，倒谱系数

1.引言

语音信号和视觉信号是人类信息和知识的主要载体，是人类进行学习和交流的重要工具。让计算机能够处理多媒体信息，以便为改善人机交互方式和丰富人们的生活已经在世界范围内受到了普遍的重视。随着近年来多媒体技术的飞速发展和许多应用领域的需求，人们对声音、图像的处理技术已有了深入的研究，但对声音和图像之间内在关系的研究还不够深入。

对视觉语音（Visual Speech）的研究正是这样一种综合考虑声音和图像的多媒体技术。视觉语音是指人们在用语言交流时所表达出的面部表情和动作，它能在一定程度上传达人们想要表达的意思，并能帮助人们加深对语言的理解。研究表明，在环境噪声较大或听话者有听力障碍的情况下，如果在给出声音信息的同时能给出一个“讲话的头”（Talking Head），即说话者面部表情和嘴部、眼部等变化情况则会大大改善人们对声音的理解。在人机交互的过程中，如果人们面对的不是单纯的文本，而是一个会说话的人物形象则使人觉得计算机界面更为友善，方便人们和计算机的交流。近几年来，对视觉语音的研究越来越受到人们的重视，已成为一个多媒体和人机交互技术研究领域相当活跃的研究方向。

无论是对唇读的研究还是对人脸图象合成的研究，首要的问题就是要建立起语音和口形的对应关系，国外已有很多学者对各种语言作了相应的研究，并已开发出商业化的产品，而国内对这一方面的研究则相对较少。在研究某种语言的基本口形时，人们一般只是根据主观的判断对各种发音的口形作了简单的分类，如 Tony Ezzat 将英语发音的口形分为 16 个基本类[1]，Tobias Ohman 将瑞典语分为 10 或 13 个基本类[2]，Cheol-Woo Jo 将朝鲜语分为 15 个基本类[3]。但我们认为这样简单的划分有两个缺点：首先，这种划分是主观的，无法确定所作的划分是否合理或是否最好；其次，对某些发音，很难用简单的一幅图象来刻划一个音节，因为它是一个连续变化的过程。在本文中我们提出一种对汉语声母韵母发音口形更客观、更准确的分类方法，即根据汉语各种发音时口形变化的过程中唇内高、唇宽、上下齿的露出程度等各个参数之间的相似度以及分类后总误差的变化曲线，将汉语声母韵母发音的口形划分为几个基本类，并可由此得出各类声母的基本口形和韵母发音的口形变化曲线。

对于由语音信号到口形参数的映射，人们也提出了多种方法，将语音信号矢量量化分类、采用神经网络或混合高斯模型分类[4]，以及采用隐马尔克夫模型[5]。在用神经网络实现由语音到口形参数转化的过程中，选取合适的网络结构和输入信号至关重要。在对发音口形参数的学习过程中，我们采用了隐含层较少但隐含结点较多的前馈神经网络结构，输入数据为语音信号的倒谱系数和平均能量，取得了较好的实验效果。

2.汉语音节发音的口形参数和分类

为了描述人们说话时的口形，我们采用了四个参数，分别是上下唇之间的高度、咀唇的宽度、上齿露出度和下齿露出度。在汉语正常说话过程中，一般语速为 3~6/秒，按每秒 25 帧计算，每个汉字约为 4~8 图象。由于声母发音时长较短，对每一个声母的发音，我们提取出具有代表意义的一幅图象来描述它；而韵母的发音占了整个汉字发音的大部分，因此对每个韵母的发音我们从整个发音过程中提取出 6 幅图象，对每一幅图象手动测得上述四个参数。这样，对每一个声母的口形我们用 4 个参数来描述；对每一个韵母我们用 24 个参数来描述。

为了能对汉语中所有的声母和韵母作一合理的分类，我们对所有可以单独发音的 55 个声母和韵母（缺韵母 ong、uo、-i 和ê）所对应的汉字作了发音录象，对 21 个声母在 4 维空间进行聚类。对于韵母，则没有必要用所有 38 去聚类，因为大多数的复合韵母的口形可由单韵母的口形组合得到，因此我们选取了 20 个韵母在 24 维空间进行口形分类。因为我们一开始并不知道需要分为多少个类，所以最初我们以每一个发音作为一个单独的类，再逐渐合并成较少的类，真到最后把所有的口形归为一类。

由 n 类合并到 n-1 类的算法如下：

(1)任取 n 类中两个类合并，构成 n-1 个类，计算其中每个类的室心；

(2)计算所有点相对所在类室心的欧氏距离，计算并记录总的类内误差和；

(3)重复(1)和(2)直到穷尽所有的 n(n-1)/2 种情况；

(4)选取总的类内误差各最小的情况构成 n-1 个类的初始划分；

(5)采用 C 均值算法进行类间调整，即对每个点作类间调整，反复迭代真到所有点相对其室心的类内误差之和达到最小，并记录分类状态和分类数为 n-1 的最小类内误差和。

重复上述算法直到合并为一个类，根据所有分类情况下的误差和与对应的分类数目，可作出分类数目与误差和的变化的对应曲线，图 1 是对汉语中 21 个声母作不同数目的分类时其类内误差和与分类数目的关系。从此曲线上我们可以看出，误差和并不是均匀变化的，我们选取分类数目的准则是在误差和尽可能小的情况下选取尽可能小的分类数目。分类数目太大，不利于我们作人脸图象合成，太小又不能准确体现说话者的口形，容易造成混淆。所以我们在上图中寻找误差明显增加前的点(拐点)来作为合适的分类数目。

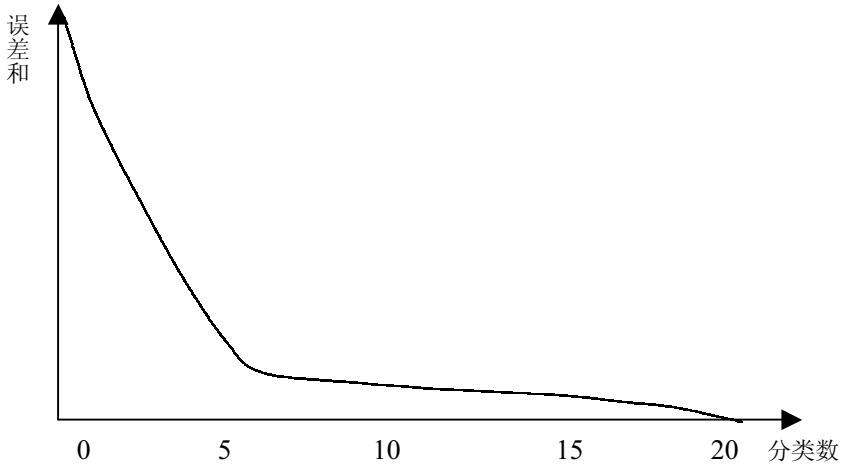


图 1 汉语声母口形分类数与总类内误差的关系

3.语音到口形参数的转换

神经网络具有自学习功能，可以描述各种变量之间复杂的映射关系。在用神经网络学习口形参数的过程中，我们用每一个网络用于学习一个口形参数，采用的神经网络结构是具有四层结构的多层感知器，学习过程中调整权值的算法采用误差反传递（BP）算法。每个结点的输出函数为 S 型非线性函数。图 2 所示为一具有两个输入、第一隐含层四个结点、第二隐含层四个结点、一个输出结点的多层感知器。

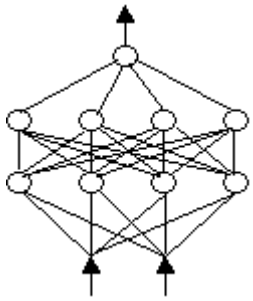


图 2 多层感知器

在学习口形参数时，我们所采用的结构输入层有 39 个结点，第一隐含层 78 个结点，第二隐含层 78 个结点，输出层一个结点。输入为相对于所测口形参数图象所在时刻前后三帧语音信号的 12 维倒谱系数及其平均能量，共计 39 个数据。学习样本为从上述汉语声母和韵母发音录象的取得的 658 幅图象(每个发音取 10~15 幅图象不等)和相应的语音波形数据。输出为一 0~1 之间的实数，代表某一个口形参数。

4.实验结果：

在口形分类实验中，我们记录了汉语中 21 个辅音的声音和脸部图像，并对其进行了分类，表 1 是对声母分类为数目分别为 6、7、8 时的分类结果。其中类内误差和为以象素点数为单位的差值平方和。从分类的结果来看，它符合人们通常发音习惯。因为韵母发音时口形有一个变化过程，对韵母的分类较为困难，也容易引入误差，表 2 所示是一种分类结果，将 21 个韵母分为 9 类。由分类结果中的类室心可得出每一个类中韵母节发音时口形各参数的变化曲线。

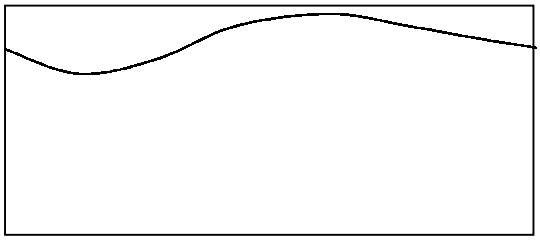
表 1 声母口形分类结果

分类数	5	6	7
类内误差和	349.08	123.00	95.92
分	b, p, m, f	B, p, m	b, p, m
	d, t, n, l	f	f
类	g, k, h	d, t, n, l	d, t, n, l
	j, q, x, z, c, s	G, k, h	g, k, h
结	zh, ch, sh, r	j, q, x, z, c, s	j, q, x
		zh, ch, sh, r	zh, ch, sh, r
果			z, c, s

表 2 韵母口形分类结果及其组合

21 个基本口形韵母的分类结果			其他韵母口形的组合
A, ai, an, ang	U		ia=i+a, ie=i+e, iao=i+ao,
ao	Ü		Iou=i+ou, ian=i+an, iang=i+ang,
o	üe,ün,iong		ua=u+a, uo=u+o, uai=u+ai,
E,ei,en,eng,er	Ou		uei=u+ei, uan=u+an, uen=u+en,
i,in, ing			Uang=u+ang, ueng(ong)=u+eng, üan=ü+an

图 3 为韵母分类后，类‘a,ai,an,ang’(实线)和‘u’(虚线)发音时的唇内高和唇宽变化曲线。其中纵坐标是以象素点数为单位的各个唇形参数幅度，横坐标为时间。



(a)唇宽　　(b)唇高

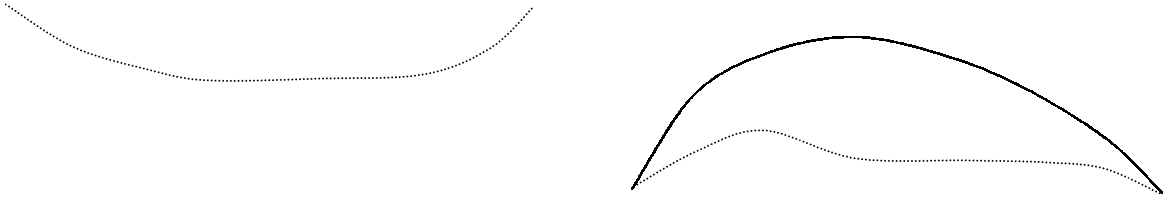


图 3 不同类口形参数的对比

在由神经网络学习口形参数的实验中，我们通过对 658 个典型数据的训练，获得网络中的各个权值。再通过非样本集中的语音信号预测相应的口形参数，图 4 是对于韵母‘uan’发音过程中唇内高预测结果与实测结果的对比，其中实线为实测结果，虚线为预测结果。从图中可以看出，虽然个别点有较大差异，但总的趋势可以反应出开口高度的变化过程。

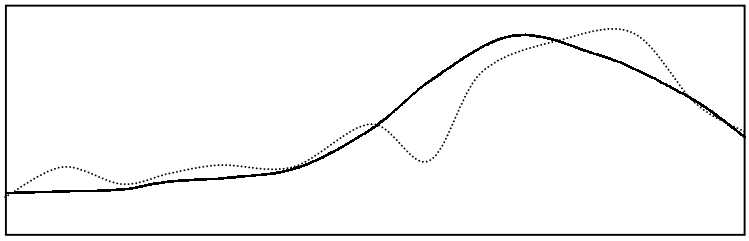


图 4 唇高预测结果与实测值的对比

5.结束语：

对发音口形的研究是合成生动逼真的人脸图象的基础，以往人们采用简单的主观判断的方法来对基本发音单元的口形进行分类，这样存在如何分类和如何选取合适的分类数的问题。本文提出一种较为客观对汉语音节发音口形分类的方法，采用分类-误差和曲线中的拐点作为合理的分类数，这样既保证了较小的误差，又得到了较少的分类数。但录象时发音的标准性和参数测量的准确性对分类结果有很大的影响，一种可能的解决方法是增加样本数目。

我们用多隐含结点的前馈神经网络实现了汉语语音到口形的预测，如何选取更为合理的网络结构还有待进一步研究。另外，通常在说话过程中，有时已有口形变化时并无声音发出，此时仅采用语音信号来预测是无法得出口形的，需要辅以别的方法。

参考文献:

[1] Tony Ezzat and Tomaso Poggio, MikeTalk: A Talking Facial Display Based on Morphing Visemes. Appears in Proceedings of the Computer Animation Conference, Philadelphia, Pennsylvania, June, 1998.

[2] Tobias Ohman, An audio-visual speech database and automatic measurements of visual speech, TMH-QPSR 1-2, P61-76, 1998.

[3] Cheol-Woo Jo, Roland Goecke, Bruce Millar, Collection of Korean audio-video speech data.

[4] Rao Ram R., Chen Tsuhan and Mersereau Russell M., Audio-to-visual Conversion for Multimedia Communication, IEEE Transcations on Industrial Electronics, v 45, n 1, Feb, 1998.

[5] Kyoung Ho Choi and Jenq-Neng Hwang, Baum-Welch hidden Markov model inversion for reliable audio-to-visual conversion, 1999 IEEE 3rd Workshop on Multimedia Signal Processing, 1999, P175-180.

Study of the Relationship Between Chinese Speech and Mouth Shape

Wang Zhiming Cai LianHong
Dep. of Computer Sci. and Tech. of Tsinghua University, Beijing, China

Abstract: In order to synthesize realistic talking-head and realize joint audio-video coding, we need to predict talker's mouth shape through speech data or text. In this paper, we proposed an efficient classing method to map basic Chinese syllable to mouth shape and obtain the profile of mouth shape for every class. In audio to visual conversion, a neural network was used to predict the mouth shape from LPC derived cepstral coefficients and average energy of speech data.

Keyword: Visual Speech, Multilayer Perceptrons, Cepstral Coefficients