基于神经网络由语音预测视位参数

王志明[1], 蔡莲红[2]

[1](北京科技大学 计算机系, 北京, 100083)
[2](清华大学 计算机系, 北京, 100084)
E-mail: wangzhiming@tsinghua.org.cn

摘 要: 语音是由多个发音器官共同作用产生的, 发音器官动作与语音之间有着内在的必然联系. 研究了利用神经网络预测视位参数中的选择语音参数、确定输入语音时域范围、优化神经网络结构等因素. 实验结果表明, 线性预测参数加短时能量优于其他语音参数, 前向协同发音较后向协同发音影响更大, 反馈对前馈神经网络的性能有所改善. 考虑到实验采用的是任意连续语流, 均方误差约为 0.0114 的实验结果还是很有吸引力的.

关 键 词: 前馈神经网络; 视位; 线性预测系数; 线谱对系数; 实倒谱系数; 反射系数; Mel 倒谱系数; 均方误差
中图分类号: TP18 **文献标识码**: A **文 章 编 号**: 1000-1220(2005)06-1083-05

Predicting Viseme Parameters from Speech Based on Neural Network

WANG Zhi-ming[1], CAI Lian-hong[2]

[1](*Department of Computer Science and Technology, University of Science and Technology, Beijing* 100083, *China*)
[2](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: Speech is produced by co-operation of all speech organs, and there are inherent relations between speech and movement of speech organs. To predict viseme parameters from speech using neural network, input speech parameters selection, time domain and structure of neural network were studied. Experiment results show that LPC coefficient plus short time energy are superior to other speech parameters, forward co-articulation is more server than backward co-articulation, and a delay feedback can improve the forward neural network performance. Considering experiments were based on unlimited vocabulary and continuous speech, the 0.0114 mean square error (MSE) is quite promising.

Key words: feed forward neural network; viseme; linear predictive coding (LPC); line spectral frequency (LSF); real cepstrum (RCEP); reflection coefficient (RC); mel frequency cepstrum coefficient (MFCC); mean square error (MSE)

1 引 言

语音是由人的各个发音器官共同作用产生的, 发音器官的动作与所处状态决定了语音的性质. 人们的发音器官有的是外界不可见的, 如肺、气管、咽喉等部分; 有些器官是外界可以看见的, 如唇、下腭等. 国际标准 MPEG-4 将发音过程中这些可视器官的所处状态定义为视位(Viseme). 人们在交流过程中, 不仅用耳朵去听声音, 而且用眼睛观察这些可视发音器官的动作, 以便获得更多的信息, 尤其是在噪声较大的环境中. 实验结果已经表明, 在强噪声环境中, 看见可视发音器官的动作, 相当于提高约 8-12dB 的语音信噪比[1]. 也就是说, 视觉信息和听觉信息的结合比任何单一信息能传达更多的信息.

然而, 视频信息的数据量远远大于相应的音频信息数据量, 这就使得在许多应用领域中, 受到各种条件的限制, 无法同时提供语音信号与视频信号, 如网络带宽的限制、存储介质容量的限制等等. 而音频与视频的内在联系告诉我们, 可以从音频数据中预测出相应的视频信息, 从而在不增加网络带宽或存储代价的情况下给人们在交流过程中提供更多的有用信息. 另外, 在可视电话、电视会议等应用环境中, 图像的主要变化正是集中在说话者的唇部, 从语音预测出口形的变化, 就可以利用音视频的交互信息来进行高效的音视频联合编码, 从而大大提高多媒体数据的压缩率.

在音视频映射的研究中, AT&T BELL 实验室的 Chen, T. 和 Rao, R. R. 等人作了长期的、大量的研究工作, 尝试了各种预测方法, 包括基于矢量量化分类的方法、基于神经网络的方法、基于混合高斯模型的方法、基于隐马尔克夫模型(HMM)的统计方法等等[2, 3, 4, 5]. Lavagetto, F. 采用了时延神经网络(Time-Delay Neural Network)的方法[6], K. H. Choi 和 William s, J. J. 等人也采用了基于隐马尔克夫模型的方法[7, 8], 陈益强等人采用了神经网络加统计约束, 由 Viterbi 寻找最优匹配的方法[9].

但以往大多数研究和实验是在小语料库上进行的, 包括少数几个单独的元音发音、孤立的数字串或英文字母发音等等. 如[2, 4, 7]中的试验语料是 4 个元音、[5]中的试验语料是 0 到 9 的 10 个英文数字读音, [6]采用了包括约 400 音节的孤立单词发音的意大利语数据库. [8][9]采用较大的语料库, 但其预测结果为图像而不是视位参数, 由于面部器官的复杂

收稿日期: 2004-05-09 基金项目: 高等学校博士学科点专项科研基金资助(20010003049); 北京科技大学校基金(2004509180)资助 作者简介: 王志明, 男, 1968 年生, 工程师; 蔡莲红, 女, 1945 年生, 教授, 博士生导师.

多变, 将所有变化归类到数十幅图像可能会造成较大的误差. 因此, 任意语料的、连续语流中的音视频映射仍是一个难题, 其主要原因是可视发音器官只是产生语音所有器官的一部分, 而且发音器官的可变性较强, 这使得语音与发音器官动作的映射关系的较为复杂, 难以提取出语音中真正与可视发音器官运动有关的特征, 也就难以在大语料库上进行音视频映射.

本文在用前馈神经网络实现音视频映射的过程中, 首先根据对大量实验结果的分析, 选择有效的语音参数作为网络输入, 包括选择合适语音参数、参数的阶数、输入语音时间范围等; 其次, 考虑到发音器官运动在时间上的连续性, 在神经网络中加入反馈连接, 并根据实验结果选择了适当的网络隐层结点数. 最终在任意语料、连续语流的音频到视位参数预测上达到了均方误差约为 0.0114 的实验结果.

2 神经网络结构

我们用于实现从语音到视位参数预测的神经网络为包含一个隐层的三层前馈神经网络, 因为三层前馈神经网络已足以实现各种复杂的非线性映射. 图 1 所示为一个输入层 4 个结点、隐层 4 个结点、输出层 2 个结点的三层前馈神经网络. 结点函数为'logsig'函数, 即每个结点的输入与输出有如下的关系:

$$y = 1/(1+\exp(-\sum_{i=1}^{N} x_i * w_i + b)) \qquad (1)$$

其中 y 为结点输出, N 为与本结点相连的输入个数, x_i 为第 i 个输入值, w_i 为第 i 个输入的权值, b 为本结点的输入偏置值. 网络的输入为归整到(-1～+1)之间的语音参数; 网络输出为归整到(0～1)之间的视位参数.

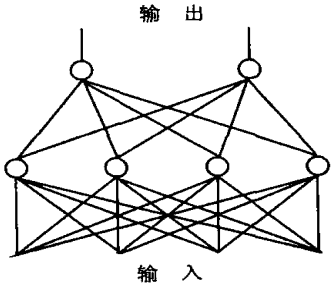

图 1 前馈神经网络示意图

为了使训练得到的网络在非训练集有更好的性能, 将整个样本集分为训练集、验证集(Validation)和测试集三部分. 在根据训练集调整网络参数的同时, 监测网络在验证集上的性能, 在发现在验证集上的性能连续五次下降时, 停止训练, 以便得到推广性能更好的网络. 测试集上的性能代表了网络对未知数据的真正性能.

3 数据准备

3.1 语料库设计

为了做到任意语料的视位参数预测, 所选取的训练数据要具有代表性. 汉语发音以音节为基本单位, 组成音节的基本单元为声韵母, 其中声母 21 个、韵母 38 个, 它们可以拼成的无调汉语音节为 412 个.

根据涵盖尽可能多音节、文本尽可能少的原则, 我们从超过 10 万字的大量现代汉语语料中, 利用程序自动选择出一个具有代表性的训练语料, 其中包括 119 个语句和短语, 共 829 个音节, 涵盖了 406 个汉语无调音节, 包括了汉语所有的声韵母. 另外, 我们录制了 10 句不包含在训练语料库中的测试语句.

3.2 实验数据的采集

虽然MPEG-4 标准仅定义了静态视位(Static Viseme), 但同时也指出不排除将来定义其它类型的视位, 我们这里所述的视位是指包括发音过程中可视器官的变化过程的广义视位概念. 用于音视频映射的视位参数包括外唇高度和宽度、上下唇突出度、下腭突出度, 以及下腭张开度. 为了能够用一种简单有效的方法测量这些参数, 我们在录像时在发音人的脸旁放置一个与人脸正面成 45 度角的镜子, 同步地记录下发音人说话过程中正面和侧面图像. 然后利用计算机图像的方法自动跟踪人脸上相关的特征点, 再根据这些特征点的位置和运动信息计算处视位参数. 在正面图中, 我们采用亮度分类后求暗点重心的方法得出鼻孔点的位置, 对唇区图像先进行可最佳区分唇色和肤色的 Finsher 变换后, 用变形模板能量最小的方法得出唇部轮廓线; 在侧面图中, 采用前景、背景分割的方法得出人脸侧面轮廓线后, 再通过计算其极值点并利用其位置信息得出各个特征点的位置. 图 2 是跟踪结果和参数测量的示意图.

图 2 视位参数的测量

我们在本文中所用的实验数据为单个女性发音人对上述语料库的发音, 总的发音长度约为 396s, 语音采样频率为 11.025KHz, 计算语音参数的帧长为 30ms, 帧移 20ms, 帧速率为 50fps(frame per second: 帧/秒); 图像分辨率 640×480, 图像帧速率为 25fps, 在实验中对视位参数数据作线性插值, 使其帧速率与语音参数相同, 达到 50fps. 总的实验样本为 19825 个, 其中训练句 17275 个样本, 测试句 2550 个样本.

4 语音特征的选择

实现一个好的模式识别系统包括两个关键步骤: 一是选择和提取最佳的特征, 二是选择有效的模式识别方法. 在已往的音视频映射研究中, 人们更多的注重采用哪一更为有效的模式识别方法, 而在一定程度上忽略了选择最佳的识别特征. 我们根据大量的实验结果选择出对某一个视位参数影响最大

的语音特征, 从而有利于下一步的视位参数预测. 语音特征的选择包括语音参数的选择、参数阶数的选择和语音时间范围的选择.

除另有说明外, 我们固定以下的测试条件: 训练目标数据为唇高参数, 神经网络的隐层结点数取输入层结点数的一半(四舍五入为整数), 从训练句的 17275 个样本中随机取 1/8 约 2160 个样本作为验证集(Validation), 其余 15115 个样本作为训练集.

衡量网络性能的参数为均方误差 MSE (Mean Square Error), 其计算公式如下:

$$mse = \frac{1}{l}\sum_{i=1}^{l}\left(\frac{y_i - f(\mathbf{x}_i)}{\Delta y}\right)^2 \qquad (2)$$

其中 $\mathbf{x}_i$ 表示第 i 个样本语音参数向量, y_i 表示第 i 个样本的某一个视位参数值, $f(\mathbf{x}_i)$ 表示神经网络预测的第 i 个样本的视位参数值, $\Delta y = \max(y) - \min(y)$ 表示参数变换幅度范围, l 表示总的样本个数.

4.1　语音参数的选择

我们用以下一些常用的语音参数作了实验: 线性预测系数 LPC (Linear Predictive Coding)、线谱频率系数 LSF (Line Spectral Frequency)、实倒谱系数 RCEP (Real Cepstrum) 和反射系数 RC (Reflection Coefficient), Mel 倒谱系数 MFCC (Mel Frequency Cepstrum Coefficient) 参数, 参数阶数取 13; 对每一种参数, 再分为单纯语音参数、语音参数加短时语音能量、语音参数加短时语音能量和过零率.

表 1　不同语音参数训练结果比较

所用参数	语音参数种类	训练集上 MSE	测试集上 MSE
语音参数	LPC	0.0187	0.0191
	LSF	0.0198	0.0200
	RCEP	0.0193	0.0197
	RC	0.0189	0.0192
	MFCC	0.0186	0.0190
语音参数 + 能量	LPC	0.0172	0.0176
	LSF	0.0183	0.0184
	RCEP	0.0183	0.0186
	RC	0.0172	0.0176
	MFCC	0.0176	0.0179
语音参数 + 能量 + 过零率	LPC	0.0201	0.0204
	LSF	0.0205	0.0208
	RCEP	0.0199	0.0209
	RC	0.0209	0.0211
	MFCC	0.0194	0.0196

因为神经网络的训练结果与初始权值有一定的关系, 我们对每一种情况进行 10 次训练, 取训练网络所能达到的最佳均方差 MSE (Mean Square Error) 作为度量其性能的参数, 实验结果如表 1 所示. 从表中可以得出以下几个结论:

(1) 各种语音参数的性能差别并不是很大, LPC 的性能相对最好;

(2) 加入短时能量参数后性能有所提高, 这是因为双唇张开的同时常常伴随时语音短时能量的增加;

(3) 加入过零率参数后性能反倒有所下降, 这可能是由于噪声的过零率常会介于清音和浊音之间, 从过零率上不易区分有无发音和唇动.

根据这一实验结果, 我们决定以 LPC 参数加语音短时能量作为神经网络的输入.

4.2　参数阶数的选择

对于选定的 LPC 参数, 需要确定其计算阶数. 从语音学上分析, LPC 参数的计算阶数一般以语音采样率(以 KHz 为单位) 加 2 较为合适. 为了确定对视位参数预测合适的 LPC 阶数, 我们在以 13 为中心的较大范围(4～22) 内进行了阶数实验, 实验中对隐层结点数固定为 10, 实验结果如图 3 所示. 图中实线为网络在训练集上达到的最佳性能(MSE), 虚线为网络在测试集上达到的最佳性能(MSE). 从图中可以看出, LPC 阶数对网络性能的影响较小. 相对来说, 在 LPC 阶数介于 8 和 13 之间时网络性能较好. 从节省计算量的角度出发, 我们认为对 11.025kHz 采样率的语音, 在进行视位参数预测时, 取语音的 LPC 阶数为 8 较为合适.

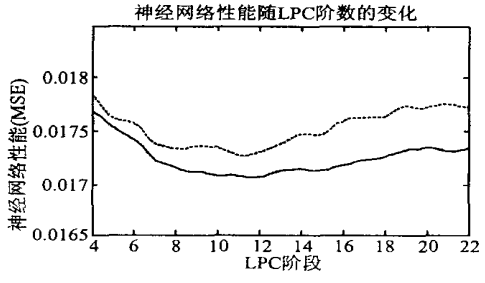

图 3　网络性能随语音 LPC 参数阶数的变化

4.3　输入语音时间范围的确定

因为发音器官运动的连续性及协同发音的影响, 每一时刻的视位参数都与其前后相邻的语音有一定的关系. 为确定影响视位参数的语音范围, 我们计算了开口高度与其邻近帧语音参数的相关性系数, 图 4 是语音参数与视位参数相关性系数的绝对值之和随时间的变化(以当前时刻为零). 从图中可以得出两个结论: (1) 但随着时间距离的增大, 这种相关性会逐渐减小; (2) 当前时刻之后(将来) 的语音比当前时刻之

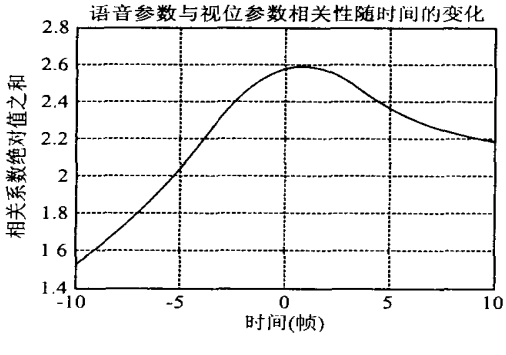

图 4　语音与视位参数相关系数的绝对值之和随时间(帧) 的变化

前(过去)的语音对口形影响更大, 也就是说前向协同发音现

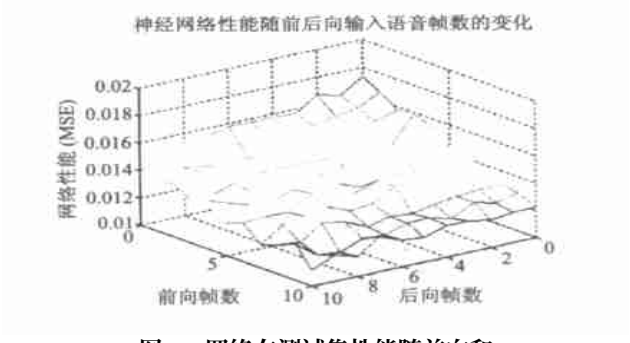

图 5 网络在测试集性能随前向和后向输入语音参数帧数的变化

象比后向协同发音现象更为严重 这同语音学中得出的结论是一致的, 即逆同化现象强于顺同化现象 对于图 4 中曲线的峰值并不在‘0’时刻, 可能是因为人们说话过程中总是先使发音器官处于适当的状态而后才发出相应的音, 在发音的同时, 发音器官的状态已开始下向一个状态过渡

为了证实以上分析的正确性, 并选择用于神经网络训练的最佳语音时间范围, 我们对输入选取不同的前向和后向帧数进行了实验, 以 20m s 为一帧, 我们对后向(预测时刻之前)计入范围从 10 到 0 帧(相当于-200m s 至 0m s)、前向(预测时刻之后)计入范围从 0 到 10 帧(相当于 0m s 至 200m s)进行了实验, 网络在训练集上的M SE 性能如图 5 所示

实验结果与图 4 基本一致, 网络性能随语音参数前向输入帧数增加的改善更为明显, 即当前时刻之后的语音比之前的语音对当前时刻的视位参数影响更大 在实际系统中, 我们可根据计算量大小和性能要求, 参考图 4 和图 5 选择适当的网络输入前向和后向帧数 实验中, 我们选择前向帧数取 4 时, 后向帧数取 2, 加上语音的短能量, 网络的输入结点确定为 (2+ 1+ 4) * (8+ 1) = 63.

5 神经网络结构的优化

在进行音视频映射过程中, 可以用同一网络预测多个视位参数, 也可以用每个网络预测一个视位参数 我们对这两种方法进行了比较, 网络的输入与隐层结点数均为 63 和 30, 结果见表 2 所示

表 2 单个网络与多个网络的性能(训练集上)比较

视位参数		唇高	唇宽	上唇突出度	下唇突出度	下腭突出度	下腭张开度	整体性能
多个网络	训练集	0. 0147	0. 0098	0. 0085	0. 0073	0. 0112	0. 0166	0. 0113
	测试集	0. 0152	0. 0103	0. 0095	0. 0080	0. 0121	0. 0170	0. 0120
单个网络	训练集	0. 0144	0. 0100	0. 0082	0. 0078	0. 0122	0. 0158	0. 0114
	测试集	0. 0149	0. 0105	0. 0088	0. 0084	0. 0127	0. 0164	0. 0119

从表中可以看出, 二者性能基本一致 从计算复杂度的角度考虑, 选一个网络预测所有六个视位参数

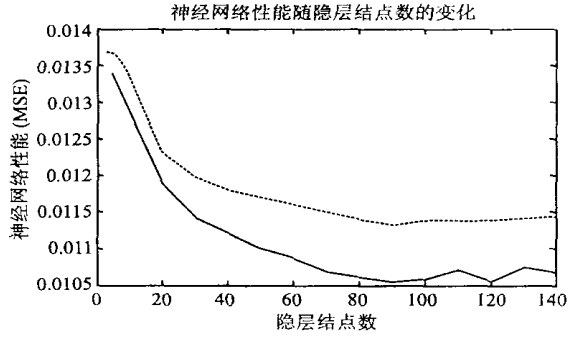

图 6 网络整体性能随网络隐层结点数的变化

从神经网络的理论上来讲, 网络隐层结点数越多, 则网络的学习能力或记忆能力越强 但网络隐层结点数越大, 需要学习的自由参数个数也越多, 也越容易产生训练样本数不足和过学习(Over Fitting)现象 我们对不同隐层结点数(5～ 140)的网络性能进行了实验, 实验结果如图 6 所示 图中实线为网络在训练集上达到的最佳性能(M SE), 虚线为网络在测试集上达到的最佳性能(M SE). 从图中可以看出, 网络性能随隐层结点数的增加在改善(同时训练时间也在增加), 但训练隐层结点数超过 30 后网络性能改变缓慢, 且测试集上性能与训练集上性能的差别变大, 说明存在过学习现象 综合网络性能和训练时间上的考虑, 我们认为取隐层结点数为 30 较为合适

另外, 人的发音器官的运动具有连续性, 因而视位参数在时间上也具有连续性 我们可以从当前时刻之前一定范围内的视位参数对当前时刻的视位参数做出一定的预测 所以, 我们在前馈神经网络的输入端增加了前几个时刻的视位参数作为反馈输入, 使得前馈神经网具有一定的反馈功能 实验结果如表 3 所示, 从表中可以看出, 加入反馈有利于改善网络性能, 但反馈数量的增加对网络性能并没有什么改善, 所以实际训练中我们取反馈数为 1.

表 3 网络整体性能随反馈时延数目的变化

反馈时延数目	0	1	2	3	4
训练集	0. 0114	0. 0105	0. 0112	0. 0111	0. 0109
测试集	0. 0119	0. 0114	0. 0117	0. 0117	0. 0116

根据以上分析, 我们最终确定的网络结构如下: 输入层为 7 帧语音的 8 阶L PC 参数和短时能量, 再加 6 个视位参数反馈共计 69 个结点; 隐层 30 个结点; 输出层 6 个结点, 分别对

应于 6 个视位参数. 图 7 是神经网络结构示意图; 图 8 是测试

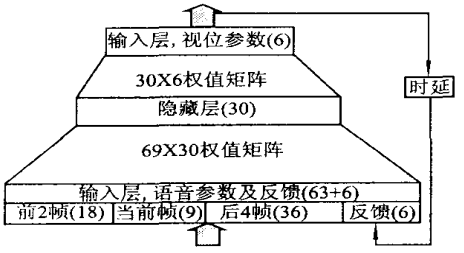

图 7　神经网络结构

集上外唇高度预测值与实测值的比较, 图中虚线为预测结果,

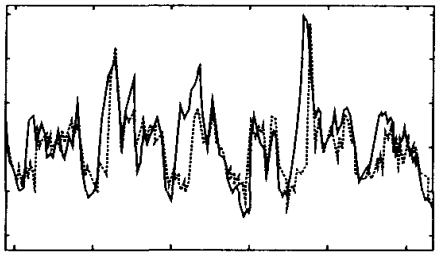

图 8　外唇高度的预测值(虚)与实测值(实)

实线为实际跟踪结果

6　结束语

本文通过理论分析和结果实验相结合的方法, 确定了在用前馈神经网络进行音视频参数映射过程中所采用的语音参数、参数阶数、语音时间范围、网络隐层结点数以及网络结构. 利用确定的语音参数和神经网络结构, 我们在任意语料、连续语流的音视位参数映射上达到了 MSE = 0.0114 的实验结果. 这一结果优于 Lavagetto 利于时延神经网络在约 400 个单词的数据库上 MSE≈0.05 的实验结果[6].

另外, 我们从实验中得知, 前协同发音(将要发出的音对当前口形的影响)比后协同发音(刚刚发出的音对当前口形的影响)更为严重, 这对于可视语音合成中处理协同发音有一定的指导意义.

在预测中我利用了当前时刻之前和之后的语音参数, 因而需要有约 80ms 的时延. 如果在实时系统中要求严格同步, 我们无法利用当前时刻之后的语音信息, 性能会略有下降.

References:

[1] Brooke N M, Scott S D, Tomlinson M J. Making talking heads and speech reading with computers [C]. IEEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication, 1996, 2/1-2/6.

[2] Rao R R, Tsuhan Chen. Cross-modal prediction in audio-visual communication [C]. In: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96), Atlanta, USA, 1996.

[3] Tsuhan Chen, Rao R R. Audio-visual interaction in multimedia communication [C]. In: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97), Munich, Germany, 1997.

[4] Rao R R. Chen Tsuhan, Mersereau Russell M. Audio-to-visual conversion for multimedia communication [J]. IEEE Trans on Industrial Electronics, 1998, 45(2): 15-22.

[5] Chen T. Audiovisual speech processing [J]. IEEE Signal Processing Magazine, 2001, 18(1): 9-21.

[6] Lavagetto F. Time-delay neural networks for estimating lip movements from speech analysis: a useful tool in audio-video synchronization [J]. IEEE Trans on Circuits and Systems for Video Technology, 1997, 7(5): 786-800.

[7] Kyoung Ho Choi, Jenq-Neng Hwang, Baum-Welch hidden markov model inversion for reliable audio-to-visual conversion [C]. In: 1999 IEEE 3rd Workshop on Multimedia Signal Processing, Piscataway, NJ, USA, 1999.

[8] Williams J J, Katsaggelos A K, Randolph M A. A hidden markov model based visual speech synthesizer [C]. In: 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00), Istanbul, Turkey, 2000.

[9] Chen Yi-qiang, Gao Wen, Wang Zhao-qi, Yang Chang-shui, Jiang Da-long. Multi-modal speech synthesis. See: Xu Ming-xing, The 6th national conference on human machine speech communication (NCHMSC6) [C]. Shenzhen, China, 2001, 163-168.

附中文参考文献:

[9] 陈益强, 高　文, 王兆其, 杨长水, 姜大龙. 多模式语音合成[A]. 见: 徐明星、第六届全国人机语音通讯学术会议论文集[C]. 深圳: 2001, 163-168.