第35卷 第17期 计 算 机 工 程 2009年9月
Vol.35 No.17 Computer Engineering September 2009
·博士论文· 文章编号：1000—3428(2009)17—0001—03 文献标识码：A 中图分类号：TP391.4

基于视频三音子的双模态语料自动选取算法

赵 晖，林成龙，唐朝京
(国防科技大学电子科学与工程学院，长沙 410073)

摘 要：为实现可视语音合成，建立符合条件的双模态语料库，提出双模态语料自动选取算法。根据视频中唇部发音特征，对已有的三音子模型归类，形成视频三音子，在其基础上从原始语料中自动选取语料，利用评估函数对原始语料中的句子打分。与其他双模态语料库相比，该语料库在覆盖率等指标上有较大改进，为实现具有真实感的可视语音合成奠定基础。

关键词：可视语音合成；双模态语料；视频三音子；评估函数

Automatic Selecting Algorithm of Bimodal Corpus Based on Visual Triphone

ZHAO Hui, LIN Cheng-long, TANG Chao-jing
(College of Electronic Science and Engineering, National University of Defence Technology, Changsha 410073)

【Abstract】 In order to realize visual speech synthesis, a satisfied bimodal database needs to be built up. This paper proposes an automatic corpus selection algorithm, according to features of lip pronunciation in video, visual triphone modal is established. Proposed algorithm automatically selects corpus from original corpus. Evaluation function is utilized to score sentences from original corpus. Compared to other bimodal databases, coverage rate, coverage efficiency and high-frequency words distribution are greatly improved, it builds a firm foundation for realistic visual speech synthesis.

【Key words】 visual speech synthesis; bimodal corpus; visual triphone; evaluation function

1 概述

近年来，可视语音合成是计算机图形学、人机交互和图形图像处理领域的一个研究热点。研究表明，在环境噪声较大或听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”，即表现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[1-3]。采用图像序列拼接的数据驱动方法是近年来研究较多的方法之一[1-2]。需要从双模态语料库中选取合适的图像实现图像序列的拼接。

因此，为了建立可视语音合成系统，必须首先建立包含有视频信息(唇部运动)和相应的音频信息的双模态语料库，大规模的双模态语料库是实现可视语音合成系统的重要基础。目前，国外已经建立的规模较大的双模态语料库包括：CMU大学的Audio-visual Speech Processing Dataset，欧洲的XM2VTSDB，IBM的ViaVoiceTM音视频库。这几个库对于以英语为基础的音视频合成和识别的研究是比较适合的。国内有中科院徐彦君等提出并建立的汉语听觉视觉双模态数据库CAVSR[4]及哈工大建立的HIT Bi-CAVDatabase[5]。它们主要是针对唇读和听觉视觉双模态语音识别的目的而建立的，规模较小，并且是以孤立音节为基础进行建库，没有考虑到连续发音中口型及语音的变化情况。

在参考其他语料库的基础上，本文提出一种基于上述因素考虑的汉语双模态连续语料的自动选取方法。

2 描述汉语连续语音的三音子结构

汉语普通话由音节连接而成，汉语的单音节由21个声母和韵母构成，韵母由9个单韵母和13个复合韵母组成(见表1)。虽然音子可以作为描述汉语普通话的最小单位，但音子在连续语流中很难以稳定形式存在。在语音学层面，同时考虑左、右相邻音子，就可以形成三音子。

表1 三音子的组成

X	Y		Z
	声母	韵母	
a, o, e, er, i, i1, i2, u, ü, n, ng，静音，21个声母	b, p, m, f, d, t, n, l, g, k, h, j, q, x, z, c, s, zh, ch, sh, r	a, o, e,i, i1, i2, u, ü,er, ai, ei, ao, ou, iu, an, en, in, un, ün,ang,eng, ing, ong, ia,ie, ua,uo,üe,iao,iou,uai,uan,uang,iang,iong	a, o, e, er, i, i1, i2, u, ü, 静音，21个声母

其中，i1对应zi, ci, si中的i；i2对应zhi, chi, shi中的i；er虽然由单韵母e和r组成，但其在发音过程中口型几乎没有发生变化，仍认为它是单韵母；X中的韵尾n与声母中的n在写法上相同，但发音位置不同，故两者有所区别。

一般来说，将音节中的声母和韵母作为中心建模单位，在考虑左右上下文变体时，只考虑其左面或右面声母或韵头的影响。这种三音子模型可以写成X-Y-Z的形式，其中，X代表左面与其相邻的声母或韵尾；Y代表声母或韵母；Z代表右面与其相邻的声母或韵头[6]。考虑到语音训练中的静音模型，可得到10种类型的三音子模型。从视觉角度上考量[7]，经过聚类分析，可将这些复合韵母作为单独的语音单位。

基金项目：国家部委预研基金资助项目
作者简介：赵 晖(1980－)，男，博士研究生，主研方向：多媒体通信，可视语音合成，网络图像安全；林成龙，硕士；唐朝京，教授、博士生导师
收稿日期：2009-04-20 **E-mail**：nudtzhaohui@sohu.com

3 考虑唇部特征的视频三音子结构

在建立双模态语料库的过程中，需要根据三音子的组合情况挑选相对应的视频信息。由于在音位和视位存在多对一的关系，因此可以根据汉语语音对应的视位将三音子精简分类，得到“视频三音子”，便于语料库视频信息的挑选。

根据文献[7]，静态视位中的唇部参数 $v_t=[x,u_0,d_0,u_1,d_1,u_2,d_2]$，利用模糊 c 均值聚类方法对声母和韵母进行分类，在视觉感知上可将声母分为 5 类：b 类，d 类，j 类，z 类和 zh 类；韵母分为 5 类：a 类，o 类，e 类，i 类和 u 类。在此基础上对表 1 中的声母、韵尾、韵母和韵头所对应的视素进行了归类，见表 2。

表 2 根据视素对汉语声母和韵母分类

类别		分类	种类
声母类		b 类	b, p, m, f
		d 类	d, t, n, l, g, k, h
		j 类	j, q, x
		z 类	z, c, s
		zh 类	zh, ch, sh, r
韵母类	X 中韵尾分类	a 类	a, an, ang
		o 类	ao, o, ong
		e 类	e, er, en,eng
		i 类	i, i1, i2, ai, ei, in, ing
		u 类	ou, iu,
	Y 中韵母分类	a 类	a,ai,ao,an,ang,ia,ua,iao,uai, uan,uang,iang
		o 类	o,ou,ong,uo,iou,iong
		e 类	e,er,ei,en,eng,ie
		i 类	i,i1,i2,in,ing
		u 类	u,ü,iu,un,ün,üe
	Z 中韵头分类	a 类	a, ai, ao, an, ang
		o 类	o, ou,ong
		e 类	e, er, ei,en,eng,
		i 类	i, i1, i2, iu, in, ing
		u 类	u,ü,ui,uo

根据唇部特征聚类结果，将韵母 er 列入 e 类，将韵母 i1, i2 列入 i 类。根据表 2 的结果，聚类简化表 1 中的声母和韵母，可归纳出视频三音子的组成，见表 3。

表 3 视频三音子的组成

X	Y		Z
	声母	韵母	
a 类，o 类，e 类，i 类，u 类，b 类，d 类，j 类，z 类，zh 类，静音	b 类，d 类，j 类，z 类，zh 类	a 类，o 类，e 类，i 类，u 类	a 类，o 类，e 类，i 类，u 类，b 类，d 类，j 类，z 类，zh 类，静音

视频三音子在不包括静音的情况下有 625 个，在包括静音的情况下有 750 个，大大简化了表 1 中的三音子类型。

4 双模态语料选取算法

4.1 双模态语料选取原则

根据语料库的建库规则和汉语双模态语料的特点，结合可视语音合成的需要，提出如下设计原则：

(1)双模态语料包括语音语料和相应的视频语料，这里视频语料指以可视语音合成为目的的唇部视频信号。

(2)考虑到汉语句子中音节内和音节间强烈的协同发音现象和视频上的连续发音现象，为了真实地反映这一现象，采用上述视频三音子模型，将三音子作为描述连续语音和视频的基本单位和基本语音现象。

(3)在可视语音合成的过程中，句子是合成的基本单位，因此，挑选语料过程中挑选的对象是真实语料中的句子。

(4)挑选的语料以文字语料为蓝本，根据文字语料录制音频语料和视频语料。

(5)采用全面覆盖原则，为避免数据稀疏，要求每个三音子至少出现一次。

(6)采用提高覆盖效率原则，用最少的语料覆盖尽可能多的语音现象。

(7)保证词语在语料中的分布律，高频词在所选语料中也应该以较高的频率出现。

(8)采用计算机自动选取方法，无需人工干预，可以根据需要调整语料选取过程中的参数来控制所选取语料的多少。

4.2 双模态语料自动选取流程

根据上述语料选取准则，借鉴文献[8-9]，设计了双模态语料的自动算法，如图 1 所示。

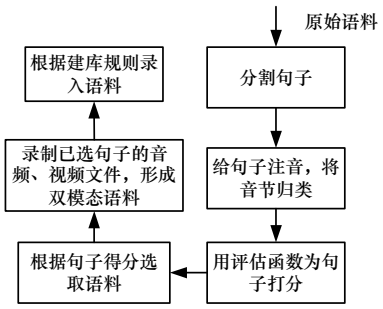

图 1 双模态语料自动选取流程

(1)分割句子时，舍弃太长或太短的句子和含有特殊符号的句子。

(2)按照汉语拼音给句子注音之后，根据声母和韵母的分类(表 2)，将三音子音节归类。

4.3 评估函数

评估函数的作用是计算原始语料中每个句子的得分，根据得分情况作为选取双模态语料的依据。设计 2 个表：一个是视频三音子表 VtriTable，存放所有视频三音子及其在已选语料中出现的次数；另一个是句子得分表 SscoreTable，存放语料中已选句子及其得分。算法如下：

(1)输入句子，并初始化该句子的得分 $SS=0$。

(2)给句子注音，求取句中的所有三音子，经归类之后得到相应的视频三音子。

(3)根据每个视频三音子对句子打分：

若该三音子在 VtriTable 中出现的次数 $n<t_1$，句子得分 $SS=SS+S_1$；

若 $t_1<n<t_2$，则 $SS=SS+S_2$；

若 $n>t_2$，则 $SS=SS+S_3$。

(4)每个视频三音子在 VtriTable 中出现的次数加 1。

(5)句子得分 $SS=SS/triNum$，$triNum$ 为该句子中视频三音子总数，将该句最终得分记入句子得分表 SscoreTable 中。

最后一步相当于求取该句子中视频三音子的平均得分，是为了避免较长的句子得到较高的分数。规定 $S_1>S_2>S_3$，$t_1<t_2$，保证了覆盖尽可能多的视频三音子，又使视频三音子不至于太稀疏。

4.4 语料选取算法

针对给定的原始语料，以视频三音子为核心，利用评估函数，双模态语料选取算法如下：

(1)输入原始语料，分割句子，设置初始值 $n=0$；

(2)用评估函数为每个句子打分，当分数低于阈值 T 时，停止打分，$n=n+1$，将句子得分记入句子得分表 SscoreTable(n) 中；

(3)根据句子得分表 SscoreTable(n)，对已打分的句子，按

照分数高低降序排序；

(4)如果 $n == N$，则转(5)，否则，重新对原始语料中的句子排序，并转到(2)；

(5)取每个 SscoreTable(n) 中得分最高的若干句子，将这些句子合并，并去掉重复的句子作为选取结果。

在实际的语料选取过程中，发现只对原始语料做一轮语料选取，选出的句子并不能真实反映语音现象的分布概率，因为处理到一定数量的句子之后，后面的句子的得分都接近于 0。所以，设置了阈值 T，并对原始语料进行了 N 轮选取，每轮选取语料之前都改变句子的录入顺序，将每轮选出的句子合并。

5 语料选取结果与分析

将选取的原始语料(40 000 句)分为 4 个文件，每个文件包含 10 000 句，对每个文件分别进行语料选取。首先从文件 1 中选取语料，在语料选取算法中，每次对句子打分之后，重新排列句子的顺序。在实际的语料选取过程中，先将文件中的句子编号 $SEN_i(i=1,2,\cdots,10\ 000)$，然后按照编号顺序选取句子，每次给句子打分后，对句子重新编号：$SEN_j(j=1,2,\cdots,10\ 000)$，再按照新编号顺序选取句子，语料选取次数 N=5, i 与 j 的关系为

$$SEN_j = SEN_{(i+1\ 0000/N)\ \mathrm{mod}\ 10\ 000} \tag{1}$$

表 4 为文件 1 选取了 N 次语料的相关信息，从表中可以看出，将每次选出的语料合并，去掉语料之间相同的句子，最终选取的句子数量较每一次有较大增长，三音子和视频三音子的覆盖率也有较大提高，分别为 79.6%和 89.9%，说明合并后的语料具有更广泛的代表性。

表 4 文件 1 所选取语料的相关信息

选取次数 N	选取的句子数	覆盖的三音子数量	三音子覆盖率(%)	覆盖的视频三音子数量	视频三音子覆盖率/(%)
1	5 982	16 548	69.7	568	75.7
2	6 015	16 785	70.7	602	80.4
3	5 873	16 823	70.8	576	76.8
4	5 932	15 930	67.1	580	77.3
5	6 167	17 532	73.8	595	79.3
合并	7 759	18 905	79.6	674	89.9

表 5 为对 4 个原始语料文件的语料选取信息，最终选取了 31 070 个句子，需要大约 160 GB 的存储空间来保存这些语料。使用高清 HDTV 进行录制，帧速率为 24 Hz，每帧图像的分辨率为 640×360。在视频采集过程中，每个采集者被要求以较均匀的速度朗读每个句子，最后经测算每个视频三音子的平均长度为 3.58 帧。采集人数为 5 人(3 男 2 女)。可以看出，将三音子归类为视频三音子之后，覆盖率有显著的提高，说明视频三音子大大降低了数据的稀疏程度。

表 5 语料文件的基本信息

文件	选取的句子数	视频文件大小/MB	单句文件平均大小/MB	视频三音子平均帧数	三音子覆盖率/(%)	视频三音子覆盖率/(%)
文件 1	7 759	40 114	5.17	3.43	79.6	89.9
文件 2	8 106	39 719	4.90	3.27	82.6	91.9
文件 3	7 258	34 984	4.82	3.84	78.5	90.4
文件 4	7 947	41 006	5.16	3.63	77.1	88.4
总计	31 070	155 823	5.02	3.58	84.1	94.0

根据 HIT Bi-CAVDatabase[5]的建库原则选取语料建立数据库，并根据实验室“十五”项目建立的语音库进行了视频录制，形成双模态语料库，将这 2 个数据库和本文算法挑选的语料库进行比较(表 6)。将每个语料库的句子分成 2 000 句一组，定义：

$$\text{覆盖效率} = \text{平均值}\left(\frac{\text{每2 000个句子覆盖的视频三音子数}}{\text{2 000个句子中的视频三音子总数}}\right) \tag{2}$$

为了描述高频词在所选语料中的分布律，统计语料中出现次数超过 30 次的视频三音子的个数及覆盖比率。比率越高，说明高频词分布越集中。

表 6 语料文件信息比较

双模态语料库	视频三音子覆盖率/(%)	覆盖效率/(%)	出现次数超过 30 次的视频三音子个数	出现次数超过 30 次的视频三音子覆盖比率/(%)
HIT Bi-CAVDatabase	80.5	0.89	423	56.4
十五语料库	85.2	0.91	580	77.3
本文语料库	94.0	1.07	591	78.8

6 结束语

本文在原有的三音子分类的基础上，将三音子合并归类，得到视频三音子，在此基础上提出了双模态语料的选取方法。本文的算法能够保证覆盖原始语料中所有语音现象，又能提高覆盖效率，保证用尽量少的语料覆盖尽可能多的语言现象。同时，语料选取算法也保证了语料中高频词的分布律。与其他指标相比，本文算法所建立的双模态语料库能够比其他语料库更好地满足上述原则。

参考文献

[1] Huang F J, Graf H P, Cosatto E. Triphone-based Unit Selection for Concatenative Visual Speech Synthesis[C]//Proc. of the Int'l Conf. on Acoustics Speech and Signal Processing. Orlando, Fl, USA: [s. n.], 2002.

[2] Cosatto E, Potamianos G, Graf H P. Audio-visual Unit Selection for the Synthesis of Photo-realistic Talking-heads[C]//Proc. of IEEE Int'l Conf. on Multimedia and Expo(Ⅱ). New York, USA: [s. n.], 2000: 619-622.

[3] Bregler C, Covell M. Video Rewrite: Driving Visual Speech with Audio[C]//Proc. of ACM Siggraph Conf. Computer Graphics. Los Angeles, USA: [s. n.], 1997.

[4] 徐彦君, 杜利民. 汉语听觉视觉双模态数据库 CAVSR1.0[J]. 声学学报, 2000, 25(1): 42-49.

[5] 洪晓鹏, 姚鸿勋, 徐铭辉. 基于句子级的唇读语料库及其切分算法[J]. 计算机工程与应用, 2005, 41(3): 174-177.

[6] 吴 华, 徐 波, 黄泰翼. 基于三音子模型的语料自动选取算法[J]. 软件学报, 2000, 11(2): 271-276.

[7] Zhao Hui, Tang Chaojing. Visual Speech Synthesis Based on Chinese Dynamic Visemes[C]//Proceedings of the 2008 IEEE International Conference on Information and Automation. Zhangjiajie, China: [s. n.], 2008.

[8] 祖漪清. 汉语连续语音数据库的语料设计[J]. 声学学报, 1999, 24(3): 236-247.

[9] 姚清耘, 刘功申, 李 翔. 基于向量空间模型的文本聚类算法[J]. 计算机工程, 2008, 34(18): 39-41.

编辑 索书志