计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
43(3):470～475, 2006

基于动态贝叶斯网络的音视频双模态说话人识别

吴志勇　　蔡莲红
(清华大学计算机科学与技术系普适计算教育部重点实验室　北京　100084)
(wuzy99@mails.tsinghua.edu.cn)

Audio-Visual Bimodal Speaker Identification Using Dynamic Bayesian Networks

Wu Zhiyong and Cai Lianhong
(*Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract　Studied in this paper is the use of dynamic Bayesian networks (DBNs) for the task of text prompt audio-visual bimodal speaker identification. The task is to determine the identity of a speaker from a temporal sequence of audio and visual observations obtained from the acoustic speech and the shape of the mouth respectively. According to the hierarchical structure of audio-visual bimodal modeling, a new DBN is constructed to describe the natural audio and visual state asynchrony as well as their conditional dependency over time. The experimental results show that the dynamic Bayesian network is a powerful and flexible methodology for representing and modeling the audio-visual correlations and the proposed DBN can improve the accuracy of audio-only speaker identification at all levels of acoustic signal-to-noise ratio (SNR) from 0 to 30dB.

Key words　biometrics; speaker identification; audio-visual bimodal modeling; fusion; dynamic Bayesian network (DBN)

摘　要　动态贝叶斯网络在描述具有多个通道的复杂随机过程方面具有优异的性能. 基于动态贝叶斯网络进行音视频双模态说话人识别的工作. 分析了音视频联合建模的层级结构,利用动态贝叶斯网络对不同层级的音视频关联关系建立模型,并基于该模型进行音视频说话人识别的实验. 通过对不同层级的建模过程及说话人识别实验的结果进行分析,结果表明,动态贝叶斯网络为描述音视频间的时序相关性和特征相关性提供了有效的建模方法,在不同语音信噪比的情况下均能提高说话人识别的性能.

关键词　生物识别;说话人识别;音视频联合建模;融合;动态贝叶斯网络

中图法分类号　TP391

1　引　　言

语音及视觉是人们之间相互交流的重要手段,也是人机交互过程中最为直接的方式. 语音发声机理的研究表明语音的发声与人脸视觉特征的变化有很大的联系. 近年来,人们已逐渐认识到语音与视觉特征之间相互关联的重要性,并进行了多方面的研究,如视觉辅助的语音识别[1～3]、音视频联合的说话人识别[4]、音视频映射[5]、虚拟人脸合成等.

关于音视频联合建模的研究,Chibelushi 等人从多个层级对音视频融合进行了分析,总结了描述音视频关联关系的多种模型,包括 HMM,multi-stream HMM,factorial HMM 等[1];Luettin 等人对音视频时序关系进行分析,并利用 multi-stream HMM 对其加以描述[2];梁路宏等人利用 coupled HMM 建立音视频关联模型,并在语音识别及说话人识别中进行实验[3,4]. 但 HMM 模型扩展性较差,模型结构改变

收稿日期:2004-11-29;**修回日期**:2005-05-27
基金项目:国家自然科学基金项目(60275014,60418012)

时,相关算法也必须随之改变;另外 HMM 模型缺乏可解释性,难以直接对音视频关联关系进行分析.

动态贝叶斯网络(dynamic Bayesian network, DBN)是近年发展起来的统计模型,能够学习变量间的概率依存关系及其随时间变化的规律,并以图的方式直观地反映这种关系,且具有很好的可扩展性和可解释性,因此适于对变量间的关联关系进行分析建模. Zweig 首先将 DBN 应用于孤立词语音识别[6],Gowdy 等人利用 DBN 进行音视频多数据流的语音识别研究[7]. 上述研究主要着眼于语音识别方面,本文尝试利用动态贝叶斯网络进行音视频双模态说话人识别的工作.

2 动态贝叶斯网络的优势

贝叶斯网络(Bayesian network)是一个有向无环图,反映了一系列变量间的概率依存关系. 而动态贝叶斯网络是贝叶斯网络在时间变化过程上的扩展,反映了一系列变量随时间变化的情况. 为方便处理,假设动态贝叶斯网络满足两个条件:①网络拓扑结构不随时间发生变化,即除去初始时刻,其余时刻的变量及其概率依存关系完全相同;②满足一阶马尔可夫条件,即给定当前时刻变量的状态后,未来时刻的状态和先前时刻的状态无关. 满足上述条件后,动态贝叶斯网络可以看做是贝叶斯网络在时间序列上的展开,如图 1 所示:

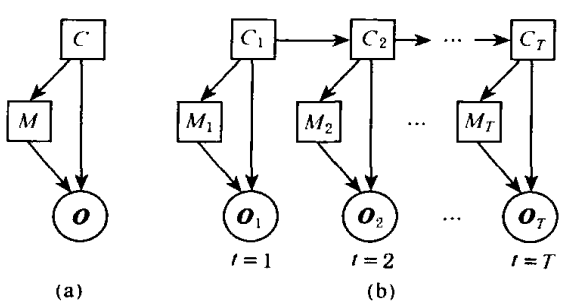


Fig. 1 Network Structure. (a) Bayesian network and (b) dynamic Bayesian network.

图 1 网络结构. (a) 贝叶斯网络;(b) 动态贝叶斯网络

考虑图 1(b)所示的动态贝叶斯网络,其描述了变量集 $X_t = \{C_t, M_t, O_t\}$ 的概率依存关系及其随时间 $t = 1, \cdots, T$ 变化的情况. 在任意时刻 t,变量 M_t 的状态由变量 C_t 决定,而 O_t 的状态则由 C_t 和 M_t 共同决定,即变量集 X_t 的联合概率分布为

$$P(X_t) = P(C_t, M_t, O_t) = P(C_t)\,P(M_t \mid C_t)\,P(O_t \mid C_t, M_t), \quad (1)$$

考虑 O_t 与 C_t 之间的条件概率分布,有

$$P(O_t \mid C_t) = \frac{P(O_t, C_t)}{P(C_t)} = \frac{\sum_m P(O_t, C_t, M_t = m)}{P(C_t)} = \frac{\sum_m P(C_t)\,P(M_t = m \mid C_t)\,P(O_t \mid C_t, M_t = m)}{P(C_t)} = \sum_m P(M_t = m \mid C_t)\,P(O_t \mid C_t, M_t = m). \quad (2)$$

在时刻 $t-1$ 和 t 之间,变量 C_t 的状态发生了转移,因此变量集 X_t 的转移概率为

$$P(X_t \mid X_{t-1}) = P(C_t \mid C_{t-1}). \quad (3)$$

可以看出,动态贝叶斯网络通过网络拓扑结构反映变量间的概率依存关系及其随时间变化的情况,其不但能够对变量所对应的不同特征之间的依存关系进行概率建模,而且对特征之间的时序关系也能很好地加以反映,因此适合于对音视频这种同时具有特征相关性和时序相关性的复杂特征进行联合建模.

除此之外,动态贝叶斯网络还可任意改变拓扑结构或增删变量以反映变量间各种不同的关联关系,而不影响训练算法本身,因此具有很好的可扩展性和灵活性;动态贝叶斯网络还具有良好的可解释性,其拓扑结构具有精确及易于理解的概率语义,通过对其进行分析可以加深对不同变量间关联关系的理解.

3 基于动态贝叶斯网络的音视频联合建模

音视频联合建模试图通过一定的策略将来自不同通道的语音及视频数据加以融合并建立模型以反映其关联关系. 本文首先对音视频联合建模的层级结构进行分析,然后针对不同层级考察音视频间的关联关系,并基于动态贝叶斯网络建立相应的音视频关联模型.

3.1 音视频联合建模的层级结构

本文将音视频联合建模的层级划分为特征级、模型级、决策级,如图 2 中椭圆形的融合模块所示:①特征级融合(feature fusion),分别对语音及视频数据进行前端处理和特征提取,然后将两者的特征参数合并为音视频联合特征参数,并对该联合特征建模和判决决策;②模型级融合(model fusion),将语音及视频的特征参数作为不同的数据源,考察它们之间存在的关联关系,为其建立联合模型;③决策级融合(decision fusion),将音视频独立,分别为其建立

模型和匹配,并将匹配结果通过决策级的融合算法进行综合.

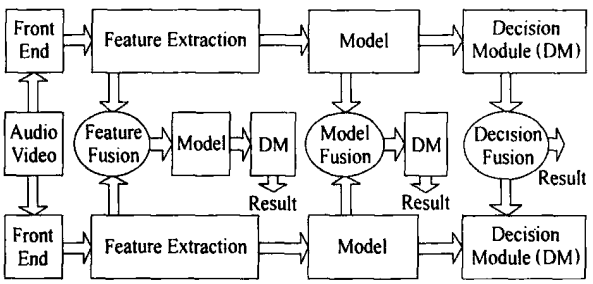


Fig. 2 Different levels for audio-visual bimodal modeling.

图 2 音视频联合建模的不同层级

不同层级的融合过程对音视频间的关联关系进行了不同程度的考虑,具有不同的建模性能.

3.2 基于动态贝叶斯网络的音视频基准模型

本文使用图 3 所示的动态贝叶斯网络结构为音频和视频特征建立基准模型,其中方框表示隐含结点、圆圈表示可观察变量对应的结点. 该网络结构描述了语句一级的模型:"单词 W"结点决定当前音视频观察序列在句中所处的单词模型;音视频"状态 C"结点取决于"单词"结点,决定观察序列在当前单词模型中所处的状态;音视频"状态转移 T"结点决定当前状态是否转移,并进一步决定"单词转移 WT"结点是否发生转移,即决定当前单词是否结束并切换到下一单词模型;上述过程不断重复,直至观察到"语句结束 EOS"为止,则整个语句的模型结束.

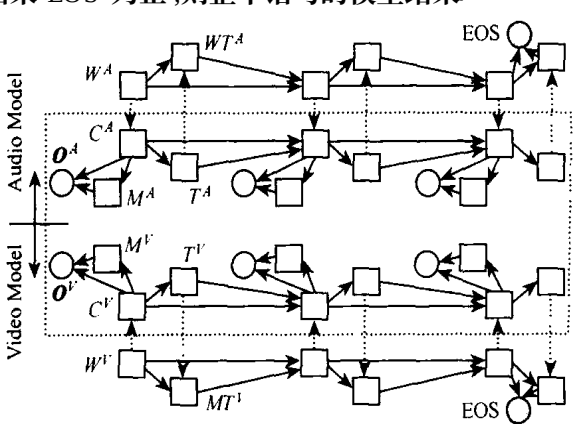


Fig. 3 Fundamental DBN model for audio and visual streams.

图 3 音频及视频基准模型的动态贝叶斯网络结构

图 3 同时给出了音视频基准模型的网络结构. 当进行音视频联合建模时,上述网络结构中,音视频状态结点 C^AC^V 及观察序列结点 O^AO^V 之间不同的组合及连接关系可以反映音视频间的不同关联特性. 本文将结合音视频联合建模的层级结构对其加以分析,即对图 3 中虚线框内所标出的部分进行分析.

3.3 语句耦合的音视频联合建模

考虑图 3 所示的动态贝叶斯网络结构,语音及视频特征分别有其独立的语句级的模型. 进行音视频联合建模时,对语音及视频序列分别进行处理,当语句结束时,对其结果进行综合,即对应于上述决策级的融合过程. 此时,语音及视频特征仅在语句一级发生关联,因此称之为语句耦合的音视频联合建模.

3.4 帧耦合的音视频联合建模

将图 3 所示的动态贝叶斯网络结构中语音及视频对应的状态结点、观察序列结点、混合数结点、状态转移结点等合并,形成新的模型结构,如图 4 所示. 其中 O_t^{AV} 为时刻 t 语音及视频特征合并(对应于观察序列结点的合并)形成的音视频联合特征参数,即动态贝叶斯网络观察结点的取值,$C_t^{AV}M_t^{AV}$ 分别为该时刻动态贝叶斯网络所处的音视频隐含状态结点和混合结点的取值. 假设在时刻 t 语音及视频的特征参数为 $O_t^s \in R^{D_s}$,其中 $s \in \{A, V\}$,D_s 为相应特征参数的维数,则音视频联合特征参数为

$$O_t^{AV} = [(O_t^A)^T, (O_t^V)^T]^T \in R^D = R^{D_A+D_V}. \quad (4)$$

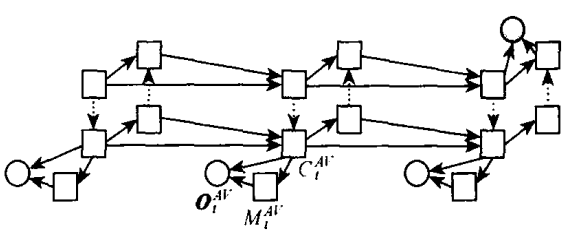


Fig. 4 DBN model for frame synchronous integration.

图 4 音视频帧耦合的动态贝叶斯网络模型

该结构中语音及视频以数据帧为单位进行特征参数的拼接,即音视频在数据帧一级完全同步,称之为帧耦合的音视频联合建模,对应于特征级融合.

3.5 状态耦合的音视频联合建模

将图 3 所示的动态贝叶斯网络结构中音视频状态结点和转移结点合并,观察序列结点、混合数结点等仍保持独立,由此形成的模型结构中,语音及视频特征分别有自己的混合结点和观察序列结点,但是共用相同的状态及转移结点,音视频在状态转移时保持同步,称之为状态耦合的音视频联合建模,如图 5所示.

音视频状态耦合的动态贝叶斯网络将音视频观察序列分开,分别考虑其在相应状态下的概率分布. 设在时刻 t 模型所处的状态为 C_t^{AV},该状态同时产生语音观察序列 O_t^A 和视频观察序列 O_t^V 的联合概率为

$$P(O_t \mid C_t^{AV}) = [P(O_t^A \mid C_t^{AV})]^{\lambda_A} \cdot$$

$$[P(O_t^V \mid C_t^{AV})]^{\lambda_V} = \prod_{s\in\{A,V\}} [P(O_t^s \mid C_t^{AV})]^{\lambda_s}, \tag{5}$$

其中 λ_A, λ_V 为语音及视频特征输出概率的融合权重,可根据语音及视频数据的质量进行设置.

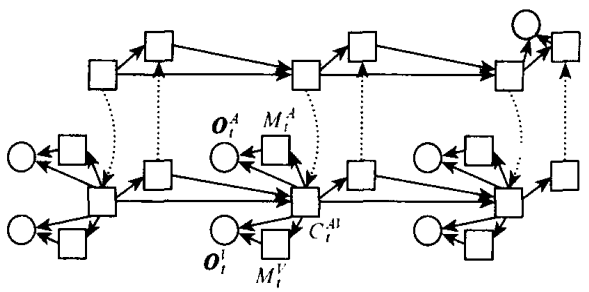


Fig. 5 DBN model for state synchronous integration.

图 5 音视频状态耦合的动态贝叶斯网络模型

3.6 特征耦合的音视频联合建模

在图 3 所示的动态贝叶斯网络结构中,考虑音视频状态转移结点之间的相关性:使得音频状态转移结点不仅决定于音频本身所处的状态,还取决于视频特征所处的状态;同样视频状态转移结点也同时取决于两者的状态. 此时,语音(或视频)状态的转移,既取决于其本身的时序关系,又决定于另一模态所处的状态,即音视频特征之间互相耦合,因此称之为特征耦合的音视频联合建模,如图 6 所示:

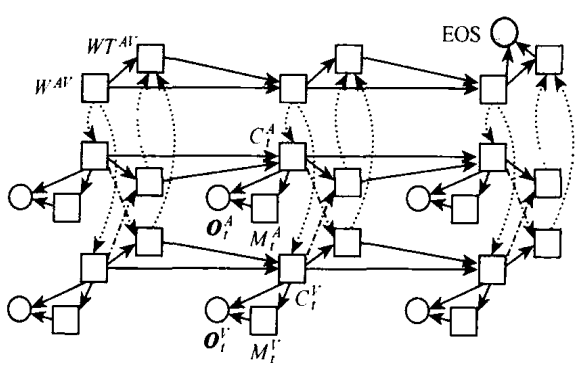


Fig. 6 DBN model for feature correlated integration.

图 6 音视频特征耦合的动态贝叶斯网络模型

4 说话人识别实验及结果分析

4.1 实验材料

实验材料包括两部分:其一来自 CMU 的音视频双模态数据库[8],该数据库采集了 7 男 3 女共 10 人的音视频数据,每人朗读 78 个单词并重复 10 次. 本实验取其中的数字部分,共 31 个单词,长约 50 s;其二为自行录制的音视频同步数据库,采集了 38 男 22 女共 60 人(年龄分布在 20～65 岁间)的数据,每人朗读长度为 2～6 个数字的连读数字串并重复 3 遍.

提取特征参数时,对音频取 13 阶 MFCC 参数和 1 阶能量参数(帧长为 25ms,帧移为 11ms)并取一阶差分,形成 28 维的语音特征参数;对视频取上唇高度、下唇高度、嘴唇宽度[8]及其一阶差分,形成 6 维的唇动特征参数. 视频的帧速率为每帧 33ms,为保证与音频帧速率相同,对视频特征参数使用线性插值法,在每两帧数据之间等间隔插入两帧新的视频特征数据,使得视频特征最终的帧速率达到和音频一致的每帧 11ms.

建模时,分别为每个数字建立动态贝叶斯网络模型:采用无跨越从左向右型的逻辑结构(即给定动态贝叶斯网络的状态数后,转移后的状态号与当前状态号相等或大于 1),语音状态数取为 5,唇动状态数取为 3,混合数均取为 3. 识别时,采用文本提示的方式进行:由提示数字串中每个数字的模型拼接形成整个数字串的模型,然后进行识别. 模型的建立和说话人识别的过程使用 FullBNT 共享工具包[9]进行.

4.2 实验结果及分析

为考察基于动态贝叶斯网络(DBN)的音视频联合建模的性能受语音噪声的影响,进行了不同语音信噪比(SNR)条件下的说话人识别的实验.

实验时,利用原始语音(SNR = 30dB)及唇动特征数据进行模型训练,然后在原始语音信号中加入高斯白噪声(white Gaussian noise)以形成不同的语音信噪比,并在不同信噪比条件下分别进行模型测试. 实验采用交叉验证(cross-validation)的方式进行:对每个用户,取该用户全部数据的 90 %进行模型训练,其余 10 %的数据用于识别,重复该过程,直至所有数据均被测试(识别)一遍. 取所有测试语句的识别结果的平均作为最终的结果,即平均识别正确率.

表 1 给出了音视频特征耦合的动态贝叶斯网络在 CMU 数据库上进行的不同语音信噪比条件下说话人识别的实验结果,其中 λ_A 表示音视频联合模型中语音所占的权重比例. 可以看出,在不同语音信噪比条件下,基于 DBN 的音视频联合模型的性能($0 < \lambda_A < 1$)比仅用语音的识别性能($\lambda_A = 1.0$)均有所提高,且信噪比较低(噪声较大)时,音视频联合模型的性能有明显提高. 另外,信噪比不同,达到最佳性能时语音的权重 λ_A 也不同,语音噪声越大说明语音质量越低,因此达到最佳识别性能时语音的权重也越低,与预期的结果一致,如表 1 中黑体部分所示.

Table 1 Feature Correlated DBN-Based Speaker Recognition Rate Dependency on the Audio Exponent for Different Values of the SNR

表 1 不同信噪比条件下特征耦合的动态贝叶斯网络的音视频说话人识别性能与语音权重的关系

SNR(dB)	λ_A							
	0.0	0.1	0.3	0.5	0.6	0.7	0.9	1.0
30	77	89	96	99	99	**100**	100	100
20	77	79	76	82	**92**	84	69	64
10	77	**79**	60	35	30	26	28	22
0	77	**60**	35	30	25	23	20	17

本文还对第 3 节中给出的反映不同音视频关联关系的动态贝叶斯网络进行了说话人识别实验. 在 CMU 数据库上进行的实验结果如表 2 所示,表 2 中给出了仅用视频特征和音频特征以及不同音视频耦合模型的说话人识别的结果,该结果为相应信噪比条件下达到的最佳平均识别正确率. 图 7 以图形方式给出了表 2 结果的直观描述. 为了考察在较大规模的数据库上本文所提出的基于动态贝叶斯网络(DBN)的说话人识别的性能,进一步在自行录制的音视频同步数据库上进行了实验,实验结果如表 3 所示.

Table 2 The Speaker Recognition Rate on the CMU Database

表 2 基于动态贝叶斯网络的说话人识别结果(CMU数据)

Models	SNR(dB)			
	30	20	10	0
Video Only	77	77	77	77
Audio Only	100	64	22	17
Sentence Synchronous	100	86	78	78
Frame Synchronous	99	85	30	20
State Synchronous	100	87	64	47
Feature Correlated	100	92	79	60

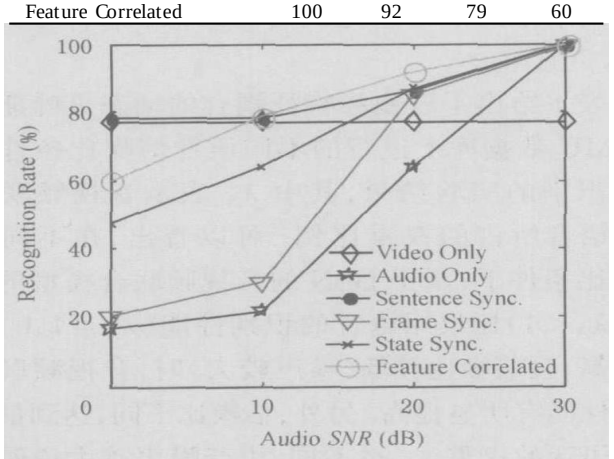


Fig. 7 The speaker recognition rate on the CMU database.

图 7 基于动态贝叶斯网络的说话人识别结果(CMU 数据)

Table 3 The Speaker Recognition Rate on Our Database

表 3 基于动态贝叶斯网络的说话人识别结果(自行录制数据)

Models	SNR(dB)			
	30	20	10	0
Video Only	74	74	74	74
Audio Only	99	59	20	15
Sentence Synchronous	100	83	76	75
Frame Synchronous	99	81	26	18
State Synchronous	100	85	61	44
Feature Correlated	100	90	75	57

分析上述实验结果可知,不同结构的动态贝叶斯网络考虑了语音及视频间的不同组合及连接关系,对音视频间的关联关系具有不同的建模性能:

(1) 帧耦合的 DBN 模型将音视频特征合并成完全同步的单一数据流,严格限制了音视频间的时序同步关系,不完全符合实际情况,因此建模性能较低.

(2) 状态耦合的 DBN 模型中音视频共用相同的状态转移序列,因此仍有较强的时序同步关系. 但与单一数据流的帧耦合模型相比,状态耦合模型分别考虑音视频序列出现的概率,并根据音视频数据的质量设置不同的权重,因此其建模性能比帧耦合模型好.

(3) 特征耦合的 DBN 模型,既考虑了音视频间的时序相关性,又考虑了其特征相关性,因此能够更准确、更有效地对音视频关联关系进行建模,从而具有更好的性能. 由表 2 可以看出,当 $SNR \geq 10$dB 时,模型同步的 DBN 具有最好的说话人识别性能.

(4) 语句耦合的 DBN 模型将音视频作为两个完全独立的模态分别处理,不考虑音视频间的相关性. 不过由于其在决策级对音视频结果进行综合考虑,因此比任何单模态都要好. 这也是表 2 中 $SNR <$ 10dB 时,语句耦合的模型反而具有更高识别性能的原因.

(5) 比较表 2 和表 3 的结果可以发现,在较大规模的数据库上说话人识别的性能有所下降,但是不同模型间的性能差别仍满足上述(1)～(4)点的结论,说明本文提出的基于动态贝叶斯网络(DBN)的音视频联合模型具有较好的建模性能及模型推广的能力.

5 结束语

本文利用动态贝叶斯网络(DBN)对音视频间不

同层级的关联关系建立模型,并基于该模型进一步进行说话人识别的实验. 结果表明,动态贝叶斯网络为描述音视频关联关系提供了有效的建模方法,在不同语音信噪比(SNR)的情况下均能提高说话人识别的性能;不同结构的动态贝叶斯网络具有不同的建模性能. 基于特征耦合的动态贝叶斯网络结构,既反映了音视频间的时序相关性,又描述了其特征相关性,因此该网络结构具有更优的建模性能,利用其进行说话人识别达到了更高的识别正确率.

针对说话人识别的任务,今后需要进一步研究如何建立更合适的说话人模型,以便更好地反映说话人的个性特征. 基于动态贝叶斯网络良好的可扩展性,可以考虑对其结构进行扩展,在其中引入更多的说话人相关的特征,如口音、语速、肤色信息等.

参 考 文 献

1 C. C. Chibelushi, F. Deravi, J. S. D. Mason. A review of speech-based bimodal recognition. IEEE Trans. Multimedia, 2002, 4(1): 23～37

2 S. Dupont, J. Luettin. Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia, 2000, 2(3): 141～151

3 A. Nefian, Luhong Liang, Xiaobo Pi, *et al*. A coupled HMM for audio-visual speech recognition. In: Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP2002). Piscataway, NJ: IEEE Press, 2002. 2013～2016

4 A. Nefian, Luhong Liang, Tieyan Fu, *et al*. A Bayesian approach to audio-visual speaker identification. In: Proc. 4th Int'l Conf. Audio- and Video-based Biometric Person Authentication (AVBPA2003). Berlin: Springer, 2003. 761～769

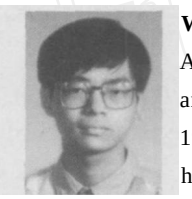
5 Wang Zhiming, Cai Lianhong, Ai Haizhou. Mouth movement prediction based on support vector regression. Journal of Computer Research and Development, 2003, 40(11): 1561～1565 (in Chinese)
(王志明, 蔡莲红, 艾海舟. 基于支持向量回归的唇动参数预测. 计算机研究与发展, 2003, 40(11): 1561～1565)

6 G. G. Zweig. Speech recognition with dynamic Bayesian networks: [Ph. D. dissertation]. Berkeley: U. C. Berkeley, 1998

7 J. N. Gowdy, A. Subramanya, C. Bartels, *et al*. DBN based multi-stream models for audio-visual speech recognition. In: Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP2004). Piscataway, NJ: IEEE Press, 2004. 993～996

8 T. Chen. Audiovisual speech processing. IEEE Trans. Signal Processing, 2001, 18(1): 9～21

9 K. Murphy. The Bayes net toolbox for Matlab. http://www.ai.mit.edu/～murphyk/Software/BNT/bnt.html, 2004-11-22



Wu Zhiyong, born in 1977. Received his B. A's and M. A's degrees in computer science and technology from Tsinghua University, in 1999 and 2001 respectively. Since 2001, he has been a Ph. D. degree candidate in computer science and technology also from Tsinghua University. His current research interests include audio-visual bimodal modeling and speech synthesis.

吴志勇,1977 年生,博士研究生,主要研究方向为音视频联合建模、语音合成.

Cai Lianhong, born in 1945. Professor and Ph. D. supervisor of Tsinghua University. Her main research interests are speech processing and synthesis, multimedia technologies, and biometrics.

蔡莲红,1945 年生,教授,博士生导师,主要研究方向为语音处理与合成、多媒体技术、生物特征识别等 (clh-dcs@tsinghua.edu.cn).

Research Background

With the development of multimedia technologies, the interaction among different media types is getting more and more attention. Research on audio-visual bimodal modeling is one of the important branches in this domain, and has significant value in audio-visual bimodal speech recognition, bimodal speaker identification, visual speech synthesis, and so on. Dynamic Bayesian networks (DBNs) are powerful and flexible methodology for representing and computing with probabilistic models of stochastic processes. We investigate the problem of audio-visual bimodal modeling with DBNs, and propose a new dynamic Bayesian network which describes the correlations between audio and visual streams as well as the asynchrony between them. The proposed DBN is then used for the task of text prompt audio-visual bimodal speaker identification, and the experimental results show that the proposed method can improve the accuracy of audio-only speaker identification at all levels of acoustic signal-to-noise ratio (SNR) from 0 to 30dB. Our work is supported by the National Science Foundation(s) of China (60275014 and 60418012).