基于发音特征的音视频说话人识别鲁棒性的研究

陈雁翔[1],刘 鸣[2]
(1.合肥工业大学计算机与信息学院,安徽合肥 230009;
2.伊利诺伊大学香槟分校电子计算机工程系,伊利诺伊州 61801)

摘 要: 人类对语音的感知是多模态的,会同时受到听觉和视觉的影响.以语音及其视觉特征的融合为研究核心,依据发音机理中揭示的音视频之间非同步关联的深层次成因,采用多个发音特征的非同步关联,去描述表面上观察到的音视频之间的非同步,提出了一个基于动态贝叶斯网络的语音与唇动联合模型,并通过音视频双模态的多层次融合,实现了说话人识别系统鲁棒性的提高.音视频双模态数据库上的实验表明了,在不同语音信噪比的条件下多层次融合均达到了更好的性能.

关键词: 发音特征;音视频;说话人识别;动态贝叶斯网络
中图分类号: TN912.34 **文献标识码:** A **文章编号:** 0372-2112 (2010) 12-2920-05

Research on Robustness of Audio-Visual Speaker Recognition Based on Articulatory Features

CHEN Yan-xiang[1], LIU Ming[2]
(1. *School of Computer Science and Information, Hefei University of Technology, Hefei, Anhui* 230009, *China*;
2. *Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Illinois* 61801, *USA*)

Abstract: Speech perception of human is bimodal because of the simultaneous audible and visible influence. This paper investigates how to fuse speech and visual speech features. From research on articulatory mechanism, the apparently observed audio-visual asynchrony is represented by asynchronous articulatory feature streams. An audio-visual model composed of speech and lip-moving is proposed based on Dynamic Bayesian Network, and then the multi-level fusion is implemented to improve the robustness of speaker recognition system. The experiment for audio-visual bimodal corpus shows that the multi-level fusion can improve the performance at all levels of acoustic signal-to-noise ratio (SNR) from 0 to 30dB.

Key words: articulatory feature; audio-visual; speaker recognition; dynamic Bayesian network

1 引言

说话人识别是语音识别研究领域中的重点之一,语音信号中除了包含语义信息外,还携带有丰富的说话人的个性信息,因此凭借语音信号中的说话人的个性特征,我们仅从电话、网络通讯、广播等中传输的声音就可以辨别和确认出说话人的具体身份.语音信号处理研究领域中的自动说话人识别就是利用计算机根据说话人的语音进行说话人身份识别的技术.

影响说话人识别系统性能的一个重要因素是训练与测试环境的失配,引起失配的主要原因之一是环境噪声.目前较有效的提高说话人识别系统的环境噪声鲁棒性的方法有多种,总的来说可以归纳为从数据、特征参数以及模型三个层面着手.从数据层面着手是指尽可能消除数据即语音信号中的噪声,如处理宽带噪声的谱减法[1].从特征参数层面着手是指提高语音特征的噪声鲁棒性,如动态参数就是一种常用的参数级抗噪方法,通过静态参数的时域差分得到的动态参数可以在一定程度上削弱平稳噪声的影响.从模型层面着手的主要方法有自适应模型修正和并行模型联合(Parallel Model Combination, PMC)[2]等.

上述三个层面的方法都是从语音单模态的角度出发,实际上,人们对语音的感知是多模态的,著名的McGurk[3]效应就说明了人类对语音的感知会同时受到听视觉的影响.进一步通过语音发声时序关系的研究发现,人类说话时音频流和视频流之间存在着异步关系,听到声音的时间基本上要比嘴形开始变化的时间平均晚大约 120ms[4].所以,建立能反映声音和唇动非同步关系的音视频双模态联合模型,将语音与视觉双模态有效地融合起来对于说话人识别鲁棒性的研究具有重要

收稿日期:2009-08-06;修回日期:2010-06-02
基金项目:国家自然科学基金(No.60672120);安徽省优秀青年科技基金(No.10040606Y09)

的意义.

由于隐马尔可夫模型(Hidden Markov Model, HMM)是当前语音识别的主流技术,目前为音视频双模态联合建模时多采用基于 HMM 的方式. Luettin 等人对音视频时序关系进行分析,并利用多流 HMM[5]在一定程度上对听视觉之间的相关性和非同步性加以描述,然而对听视觉非同步关联关系的建模停留在音素级,研究实验证明,由于协同发音现象的普遍存在使得听视觉间的非同步关联已经超过了音素边界. 另外一种常用的乘积 HMM[5]带来了状态空间过大、计算量增加等问题. Stephen 等人利用耦合 HMM[6]在语音识别中进行了实验,不同信噪比下均达到不错的效果. 但是,我们注意到,基于 HMM 的方式对于表述音视频双模态融合这样复杂的问题有致命的弱点,这主要表现在:HMM 模型的扩展性较差,模型结构改变时,相关算法也必须随之改变;并且,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.

所以,本文首先利用动态贝叶斯网络(Dynamic Bayesian Network, DBN)[7]建立音视频联合模型,因 DBN 具有可扩展性和解释性,适于对特征之间关联关系进行描述. 其次,人类发音机理的研究揭示了音视频之间非同步关联的深层次成因,即表面上观察到的语音与唇动特征的非同步本质上是多个发音特征在发音过程中的非同步,因此,我们建立了基于多个发音特征流的非同步关联的音视频联合模型,把多个发音特征作为多个隐含的状态变量,输出的语音、唇动特征观察值概率由各个发音特征状态变量共同作用;允许各个发音特征流之间存在非同步的关联,并对非同步的程度加以约束. 音视频双模态数据库上的实验表明了该模型能提高说话人识别系统的噪声鲁棒性.

2　基于非同步发音特征流的音视频联合模型

2.1　基于动态贝叶斯网络的音视频基准模型

动态贝叶斯网络能以图的方式直观地反映变量间的概率依存关系及其随时间变化的规律,非常适合对时间序列进行建模. 它还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,因为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,并且,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,因此适于对音视频间的关联关系进行分析建模. 本文使用图 1 所示的动态贝叶斯网络结构为音视频特征建立基准模型.

图 1 所示为连续的某两帧,由此得到联合条件概率分布如下:

$$\begin{aligned} p(\Lambda_t \mid \Lambda_{t-1}) = {} & p(\boldsymbol{x}_t \mid q_t)\, p(\boldsymbol{y}_t \mid q_t)\, p(q_t \mid \varphi_t, w_t)\, p(qinc_t \mid q_t) \\ & \cdot p(\varphi_t \mid \varphi_{t-1}, qinc_{t-1}, winc_{t-1}) \\ & \cdot p(winc_t \mid qinc_t, \varphi_t, w_t) \\ & \cdot p(w_t \mid w_{t-1}, winc_{t-1}) \end{aligned} \tag{1}$$

其中:$p(\Lambda_t \mid \Lambda_{t-1})$表示给定 $t-1$ 帧的所有变量,产生 t 帧的所有变量的条件概率.

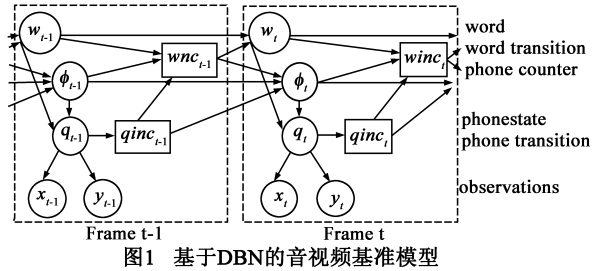

图1　基于DBN的音视频基准模型

在图 1 中,观察值(observations)向量对应的节点是可观测节点,其余的均为隐含节点. 以 t 帧为例加以说明:$\boldsymbol{x}_t$ 和 $\boldsymbol{y}_t$ 分别表示音频和视频的特征向量,类似于隐马尔科夫模型,概率分布 $p(\boldsymbol{x}_t \mid q_t)$和 $p(\boldsymbol{y}_t \mid q_t)$采用多个高斯分布的加权和来描述. q_t 是音素的状态(phonestate). $qinc_t$(phonestate transition)表示音素状态的转移,当发生转移时其值为 1,否则为 0. ϕ_t(phone counter)表示音素在词中的位置,初始帧时 ϕ_t 为 1;在其它帧,当一个词结束发生词间转移时($winc_{t-1}=1$),其值复位为 1;当在一个词的内部且发生了音素状态转移时($winc_{t-1}=0$, $qinc_{t-1}=1$),则 $\phi_t=\phi_{t-1}+1$;当既没有词间转移也没有音素状态转移时,$\phi_t=\phi_{t-1}$. $winc_t$(word transition)表示词的转移,当发生了音素状态转移($qinc_{t-1}=1$),且 ϕ_t 的值为所在词包含的音素的状态数之和时,则说明有词间转移,$winc_t$ 的值赋为 1,否则为 0. w_t 表示词(word).

语音和唇动特征在模型中可共用相同的状态转移序列(如图 1 所示),即它们的关联关系是同步的,但语音发声的时序关系表明音频和视频之间存在着异步关联,因而,一个更好的选择是采用非同步关联的方式.

2.2　发音特征流的非同步关联

发音音位学的研究工作指出,在发音过程中,图 2 所示的 8 个声道变量中的每一个都可以用对应的 gestures 值来描述,如 LIP-LOC(Lip Location)的值有:Protruded(伸出、突出),Labial(唇音的,表示嘴唇通常所处的位置),Dental(齿音的,表示唇齿相接时的位置);LIP-OP(Lip Opening)的值有:Closed

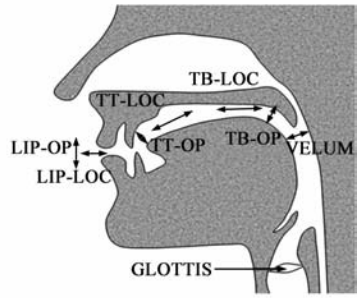

图2　发音过程中声道变量示意图

(闭合),Critical(临界闭合),Narrow(张开度小),Wide(张开度大).

对 8 个声道变量进行合理地简化,组合成 3 个,即:LIP-LOC 和 LIP-OP 组合成 Lip;TT-LOC,TT-OP,TB-LOC 和 TB-OP(TT 指 Tongue Tip,TB 指 Tongue Body)组合成 Tongue;GLOTTIS 和 VELUM 组合成 Glottis,并分别用字母 L,T 和 G 来表示,则 L 是描述嘴唇所处位置及张开度的发音特征,T 是描述舌尖和舌体的发音特征,G 是描述软腭和声门的发音特征.根据 $\{L, T, G\}$ 对应的 gestures 值,可以建立各音素与发音特征值之间的对应关系表[8].

在发音过程中,L、T 和 G 发音特征之间可以是非同步的.如在发"three"音的过程中,说话人的舌尖和嘴唇已分别处于词"three"的前两个音素(/θ/和/r/)的状态,而此时的静音说明声门处于"Silent"值.图 3 对此做了解释,在听到"three"音前的某一静音时刻,三个发音特征分别对应不同的音素:G 对应静音(对应值为"Silent"),T 对应第 1 个音素/θ/(对应值为"Dental Critical"),L 对应第 2 个音素/r/(对应值为"Round").所以,听到声音的时间要比嘴形开始变化的时间晚.

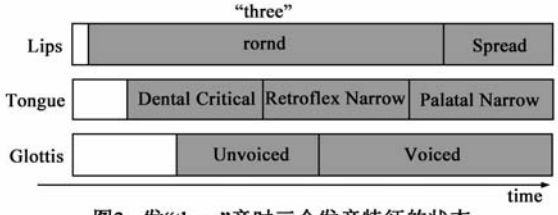

图3 发"three"音时三个发音特征的状态

上述发音机理的研究表明了:表面上观察到的音视频之间的非同步本质上是多个发音特征在发音过程中的非同步.以此为出发点,我们的研究思路是:不再如传统的 HMM 那样用音素作为隐含的状态变量,取而代之的是发音特征,把发音特征作为隐含的状态变量,允许各个发音特征流之间存在非同步的关联,并对非同步的程度加以约束.

2.3 基于动态贝叶斯网络的发音特征模型(AFM)

我们用图 4 所示的动态贝叶斯网络结构来表达发音特征模型(Articulatory Feature Model,AFM).其中,三个发音特征分别作为三个隐含的状态变量 l_t, t_t, g_t,各个发音特征流之间是非同步的关联,即有各自的状态转移 $linc_t, tinc_t, ginc_t$,同时利用耦合变量值的概率分布实现对非同步程度的约束,在图中,由于篇幅限制只画出了三个耦合变量中的两个:$\delta_t = |\lambda_t - \tau_t|$ 和 $\varepsilon_t = |\tau_t - \gamma_t|$(另一个是 $\theta_t = |\lambda_t - \gamma_t|$).以 δ_t 为例,它表示 L 和 T 发音特征流中当前所处的状态值的绝对差,反映了非同步的程度;耦合变量的值服从一定的概率分布,即有 $P(\delta_t = a \mid a > \delta_{\text{Max}}) = 0$,其中 δ_{Max} 是允许的最大的非同步程度,我们取为 2,此式说明 δ_t 耦合变量的值只有是 0、1、2 时才有非零的概率,使得发音特征流之间的非同步不能超过允许的范围,由此控制 $linc_t$ 和 $tinc_t$,控制相应特征的状态转移概率.

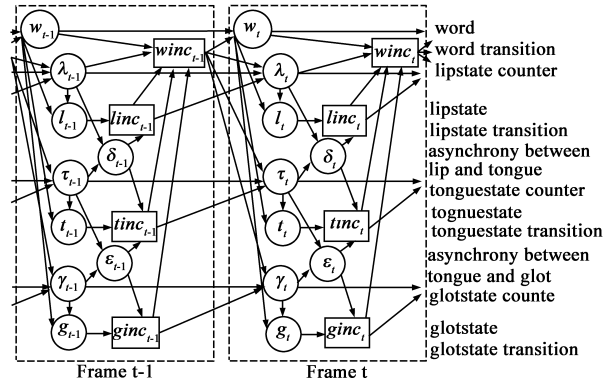

图4 基于DBN的发音特征模型

同样由于篇幅限制,图中未画出观察值变量 $\boldsymbol{x}_t$ 和 $\boldsymbol{y}_t$,但它们的观察值输出概率均采用由各个发音特征状态变量共同作用的方式:对于观察值变量 $\boldsymbol{x}_t$,输出概率 $P(\boldsymbol{x}_t \mid l_t, t_t, g_t)$ 表示语音特征 $\boldsymbol{x}_t$ 在三个发音特征状态变量共同作用下的条件概率;对于观察值变量 $\boldsymbol{y}_t$,输出概率 $P(\boldsymbol{y}_t \mid l_t, t_t)$ 表示唇动特征 $\boldsymbol{y}_t$ 在 L 和 T 发音特征状态变量共同作用下的条件概率,我们认为 G 发音特征和视觉参数无关.上述两个条件概率都用多个高斯分量的加权和来表示.

2.4 音视频双模态的多层次融合

在上述模型级融合的基础上,我们进行决策级的融合,从而实现如图 5 所示的语音与视觉特征的多层次融合.决策级融合中将匹配结果即概率评分通过融合算法进行综合,利用了多个特征之间的互补性,且遵循低信噪比时唇动特征由于抗噪性强应具有较高可靠度和权重的准则,以期利用多个特征的综合结果来提高系统的性能.

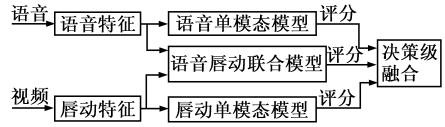

图5 多层次融合的系统框图

最终输出的联合概率密度如下式所示:

$$P(\boldsymbol{O}_A, \boldsymbol{O}_V \mid M_A, M_V, M_{AV}) = [P(\boldsymbol{O}_A \mid M_A)]^{\lambda_A} \cdot [P(\boldsymbol{O}_V \mid M_V)]^{\lambda_V} [P(\boldsymbol{O}_A, \boldsymbol{O}_V \mid M_{AV})]^{\lambda_{AV}} \quad (2)$$

其中:$P(\boldsymbol{O}_A \mid M_A)$ 是语音模型 M_A 产生语音特征矢量 $\boldsymbol{O}_A$ 的条件概率,其余类推,M_V 和 M_{AV} 分别对应唇动模型和语音与唇动联合模型.指数 λ_A,λ_V 和 λ_{AV} 反映了融合时的权重,满足 $\lambda_A + \lambda_V + \lambda_{AV} = 1$,$\lambda_A = \lambda_{AV}, \lambda_A, \lambda_V, \lambda_{AV} \geqslant 0$ 的约束.

权重的评估应随着背景环境的变化而变化,在不

同的语音信噪比条件下,各模型对最终识别性能的贡献是不同的,当信噪比较低即语音质量较差时,唇动模型由于抗噪性强,应具有更高的权重,而赋予与语音相关的模型较低的权重.

3　说话人识别实验及结果分析

3.1　实验数据库及特征提取

实验所用数据来自 CMU 的音视频双模态数据库,其中采集了 7 男 3 女共 10 人的音视频数据,每人朗读 78 个单词(包含数字、星期、月份等),并重复 10 次.我们取其中的数字部分,共 31 个单词.

提取特征参数时,对音频取 13 阶 MFCC 参数和 1 阶能量参数并取一阶差分,形成 28 维的语音特征参数;对视频取上唇高度、下唇高度、嘴唇宽度及其一阶差分,形成 6 维的唇动特征参数.由于语音的帧长为 25ms,帧移为 11ms,而视频的帧速率为每秒 30 帧,所以通过在相邻视频帧之间进行插值实现升采样,使视频的帧速率达到与音频的一致.

建模时,分别为每个说话人建立各个数字的模型.对于 CHMM 模型(参考文献[6]的基础上用动态贝叶斯网络实现),语音状态数取为 5,唇动状态数取为 3,混合度数均为 3;对于本文提出的 AFM 模型,L,T,G 发音特征的状态数均取为 3,混合度数均为 2.识别时采用基于数字串的文本提示的方式进行,即要求说话人按提示的数字串发音,将其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.模型的训练和识别的过程均使用 GMTK(Graphical Models Toolkit)工具包[9]进行.

3.2　实验结果及分析

为考察基于动态贝叶斯网络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以形成不同的语音信噪比,并在不同的信噪比条件下进行识别,测试模型的噪声鲁棒性.实验采用交叉验证的方式进行,对每个说话人,取其全部数据的 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一遍,并取所有测试语句的识别结果的平均作为最终的结果.

表 1　不同信噪比条件下说话人识别的正确率(%)

Audio signal-to-noise ratio(SNR)	30dB	20dB	10dB	0dB
Audio-only	100	64	22	17
Video-only	77	77	77	77
CHMM	100	90	78	60
AFM	100	92	80	65
Multi-level fusion	100	93	82	79

分析表 1 的实验结果可知:

(1)Audio-only 的单模态音频系统在低信噪比时根本无法工作;Video-only 的单模态视频系统的性能与信噪比无关;而音视频双模态系统的性能在所有带噪环境下均优于单模态音频系统.

(2)AFM 模型比 CHMM 模型具有更高的识别正确率,尤其在低信噪比时,这是由于前者揭示了音视频之间非同步关联的深层次成因,把语音、唇动特征之间的非同步关联关系更准确地描述出来.

(3)由于训练和测试环境的失配,在 SNR 小于 10dB 时,无论 AFM 还是 CHMM 模型,它们的性能均不及 Video-only 单模态视频系统.

(4)多层次融合解决了(3)的问题,它通过对音视频特征进行综合考虑,且遵循低信噪比时唇动特征由于抗噪性强应具有较高权重的准则,从而使得 Audio-only 和 Video-only 成为 AFM 模型的有效补充,增强了说话人识别系统的鲁棒性.权重 λ_A 的值在 30dB、20dB、10dB 和 0dB 条件下分别为:0.4、0.3、0.1、0.01,满足 $\lambda_A + \lambda_V + \lambda_{AV} = 1$,$\lambda_A = \lambda_{AV}$,$\lambda_A, \lambda_V, \lambda_{AV} \geqslant 0$ 的约束.

4　结论

人类对语音的感知会同时受到听觉和视觉的影响,本文提出了一个利用音视频双模态提高说话人识别鲁棒性的新途径.这一途径的核心思想是采用基于多个发音特征流的非同步关联的描述,来建立语音和唇动特征相互耦合的联合模型;并通过音视频双模态的多层次融合,来实现对语音及其视觉特征的综合考虑.实验结果表明,多层次融合在不同语音信噪比的条件下均达到了更好的性能.

针对说话人识别的任务,今后需要进一步研究能更好地反映说话人个性的特征,利用动态贝叶斯网络良好的可扩展性,通过对其结构进行扩展,在其中引入更多的说话人相关的特征,如脸像特征、语速等.

参考文献:

[1] Weber F, Peskin B, et al. Speaker recognition on single and multi-speaker data[J]. Digital Signal Processing, 2000, 10(1): 75 – 92.

[2] Nakagawa S, Zhang W, Takahashi M. Text independent speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM[A]. Proceedings of the International Conference on Acoustic, Speech and Signal Processing (ICASSP)[C]. Montreal, Canada: Institute of Electrical and Electronics Engineers Inc, 2004. 81 – 84.

[3] McGurk H, MacDonald J. Hearing lips and seeing voices[J]. Nature, 1976, 264(5588): 746 – 748.

[4] Brady K, Brandstein M. An evaluation of audio-visual person recognition on the XM2VTS corpus using the Lausanne protocols[A]. Proceedings of the International Conference on Acoustic, Speech and Signal Processing (ICASSP)[C]. Hawai'i, USA: Institute of Electrical and Electronics Engineers Inc, 2007.237 – 240.

[5] Chen T. Audio-visual speech processing[J]. IEEE Transactions on Signal Processing, 2001, 18(1): 9 – 21.

[6] Chu S M, Huang T S. Multi-model sensory fusion with application to audio-visual speech recognition[A]. Proceedings of European Conference on Speech Communication and Technology (Eurospeech)[C]. Aalborg, Denmark: European Association for Signal Processing, 2001.109 – 112.

[7] Gowdy J N, Subramanya A, Bartels C, Bilmes J. DBN based multi-stream models for audio-visual speech recognition[A]. Proceedings of the International Conference on Acoustic, Speech and Signal Processing (ICASSP)[C]. Montreal, Canada: Institute of Electrical and Electronics Engineers Inc, 2004. 993 – 996.

[8] Browman C P, Goldstein L. Articulatory phonology: An overview[J]. Phonetica, 1992, 49: 155 – 180.

[9] Bilmes J, Zweig G. The graphical models toolkit: An open source software system for speech and time-series processing [A]. Proceedings of the International Conference on Acoustic, Speech and Signal Processing (ICASSP)[C]. Florida, USA: Institute of Electrical and Electronics Engineers Inc, 2002.3916 – 3919.

作者简介:

陈雁翔 女, 1972 年生于安徽合肥. 2004 年毕业于中国科学技术大学电子科学与技术系, 获工学博士学位. 2006 年至 2008 年赴美国伊利诺伊大学香槟分校电子计算机工程系做访问学者. 研究方向为语音信号处理、模式识别.

E-mail: hfutchenyx716@hotmail.com