基于多流多状态动态贝叶斯网络的音视频连续语音识别

吕国云[①]　蒋冬梅[①]　张艳宁[①]　赵荣椿[①]　H Sahli[②]　Ilse Ravyse[②]　W Verhelst[②]
[①](西北工业大学计算机学院　西安　710072)
[②](布鲁塞尔自由大学电子与信息处理系　布鲁塞尔　B-1050　比利时)

摘　要：语音和唇部运动的异步性是多模态融合语音识别的关键问题，该文首先引入一个多流异步动态贝叶斯网络(MS-ADBN)模型，在词的级别上描述了音频流和视频流的异步性，音视频流都采用了词-音素的层次结构。而多流多状态异步 DBN(MM-ADBN)模型是 MS-ADBN 模型的扩展，音视频流都采用了词-音素-状态的层次结构。本质上，MS-ADBN 是一个整词模型，而 MM-ADBN 模型是一个音素模型，适用于大词汇量连续语音识别。实验结果表明：基于连续音视频数据库，在纯净语音环境下，MM-ADBN 比 MS-ADBN 模型和多流 HMM 识别率分别提高 35.91%和 9.97%。

关键词：语音识别；动态贝叶斯网络；音视频；多流异步
中图分类号：TP391.42　　**文献标识码**：A　　**文章编号**：1009-5896(2008)12-2906-06

DBN Based Multi-stream Multi-states Model for Continue Audio-Visual Speech Recognition

Lü Guo-yun[①]　Jiang Dong-mei[①]　Zhang Yan-ning[①]　Zhao Rong-chun[①]
H Sahli[②]　Ilse Ravyse[②]　W Verhelst[②]
[①](*Northwestern Polytechnical University*, *School of Computer Science*, *Xi'an* 710072, *China*)
[②](*Vrije Universiteit Brussel*, *Department ETRO*, *Brussel B*-1050, *Belgium*)

Abstract: Asynchrony of speech and lip motion is a key issue of multi-model fusion Audio-Visual Speech Recognition (AVSR). In this paper, a Multi-Stream Asynchrony Dynamic Bayesian Network (MS-ADBN) model is introduced, which looses the asynchrony of audio and visual streams to the word level, and both in audio stream and in visual stream, word-phone topology structure is used. However, Multi-stream Multi-states Asynchrony DBN (MM-ADBN) model is an augmentation of Multi-Stream DBN (MS-ADBN) model, is proposed for large vocabulary AVSR, which adopts word-phone-state topology structure in both audio stream and visual stream. In essential, MS-ADBN model is a word model, and while MM-ADBN model is a phone model whose recognition basic units are phones. The experiments are done on small vocabulary and large vocabulary audio-visual database, the results show that: for large vocabulary audio-visual database, comparing with MS-ADBN model and MSHMM, in clean speech environment, the improvements of 35.91 and 9.97% are obtained for MM-ADBN model respectively, which show the asynchrony description is important for AVSR systems.

Key words: Speech recognition; Dynamic Bayesian Network (DBN); Audio-visual; Multi-stream asynchrony

1　引言

多模态音视频语音识别是近年来语音信号处理新的研究热点[1,2]。结合人说话时的唇部视觉特征，可以提高噪声环境下语音识别的鲁棒性。然而心理声学研究和音视频融合模型的实验结果表明：虽然人的唇部视觉运动和声音是相关的，但是并不同步，唇部运动先于语音信号大约 120ms 左右[2]，任何音视频联合建模的语音识别系统都应该尽可能考虑这个事实。

对于音视频模型融合的语音识别研究，Potamianos, Nefian 等人对状态同步/异步多流 HMM, 乘积 HMM, coupled HMM, factorial HMM 等多流 HMM 进行了分析和识别实验[1−3]。虽然多流 HMM 能在一定程度上(状态，音素，音节)反映音视频流的异步性，但是对中大词汇量的音视频语音识别，多流 HMM 仅能采用音素基元来建立模型，限制音视频流的异步性在音素边界，并不能充分描述音视频流之间的异步关系。

利用动态贝叶斯网络(Dynamic Bayesian Network, DBN)模型进行语音识别研究是近年来的一个研究热点，DBN 能够描述变量之间的概率依赖关系及随时间变化的规律，适合对复杂的变量关系进行建模。Bilmes, Zweig 等人采

2007-06-11 收到，2007-11-27 改回
中国科技部与比利时弗拉芒大区科技合作项目([2004] 487)和西北工业大学英才培养计划项目(04XD0102)资助课题

用 DBN 模型来研究小词汇量的语音识别[4−6]。Gowdy 建立了多流 DBN 模型[5]，通过词转移概率的发生迫使音频流和视频流在词节点同步，然而关于音视频流如何决定词转移的发生，却没有给出具体的描述。Bilmes 给出了一个通用的多流异步的 DBN 模型结构[6]，描述了词转移概率和音视频流相关节点之间的条件依赖关系，但没有给出实验结果。同时上述模型在本质上都是词模型，识别基元是词，仅适合于小词汇量的音视频语音识别任务。

本文基于 Bilmes 提出的多流 DBN 模型结构，首先把原模型中的词-状态的构成形式更改为词-音素的构成形式，这样由于音素由多个词共享，减少了训练参数，本文称之为多流异步 DBN (MS-ADBN)模型。同时在 MS-ADBN 模型的音视频流的拓扑结构中各增加了一个隐含的状态节点层，构成了一个新颖的多流多状态异步 DBN (Multi-stream Multi-states Asynchrony DBN，MM-ADBN)模型。音频流和视频流都采用了词-音素-状态的层次结构，识别基元为音素，模型不但描述了词的动态发音过程，而且描述了音素的动态发音过程，适合对大词汇量进行音视频连续语音识别。

2　多流异步 DBN 模型

2.1　MS-ADBN 模型

图 1 描述了 MS-ADBN 模型的语音识别结构，在这个模型中，词(word)和词转移节点(word transition)位于模型的上方，在两个词之间，音频流和视频流各自有独立的音素(phone), 音素位置(phone position), 观测向量，音素转移概率以及节点变量之间的条件依赖关系。在音频流和视频流，每个词都是由它的对应音素构成，相当于具有对词节点的两个独立的描述，而每个音素和观测向量直接联系。当词转移发生时，音频流和视频流中的音素位置节点都被强行复位而迫使音频视频流在词节点同步。而词转移概率由音频流和视频流的节点变量共同确定，对于确定的词，只有当音频流和视频流中的音素为在该词中的最后一个音素，并且两流的音素转移都同时发生时，词转移才会发生。然而 MS-ADBN 模型本质上是一个词模型，仅适合于小词汇量数据库的音视频语音识别。

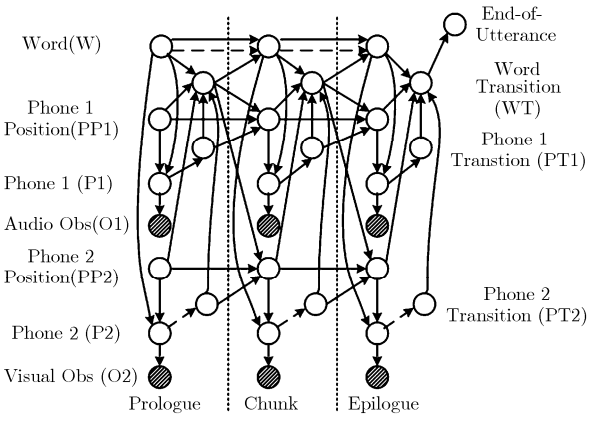

图 1　MS-ADBN 音视频连续语音识别模型

2.2 MM-ADBN 模型

为了能对大词汇量进行音视频语音识别，应该采用更小的识别基元-音素，基于 MS-ADBN 模型，本文在音视频流的拓扑结构中都增加了一个隐含的状态节点层，构成 MM-ADBN 模型，见图 2 所示，在音频流和视频流，每个词由它的对应音素构成，而音素由固定个数的状态描述，状态和观测向量相联系，它的识别基元是音素，可以满足大词汇量数据库音视频语音识别的任务。

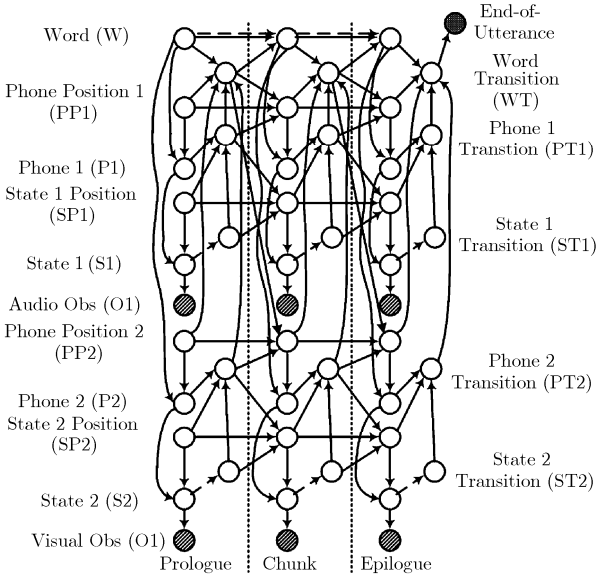

图 2　MM-ADBN 音视频连续语音识别模型

图 2 中圆括号内为对应节点变量的简称，W 为词节点，WT 为词转移概率，P1 和 P2 为音素节点，PP1 和 PP2 表示音素在词中的位置，PT1 和 PT2 为音素转移概率，S1 和 S2 为状态节点，SP1 和 SP2 为状态在音素中的位置，ST1 和 ST2 为状态转移概率，$\boldsymbol{O}1$ 为音频特征观测向量，$\boldsymbol{O}2$ 为视频特征观测向量。

下面详细描述了主要节点变量及其条件概率分布(Conditional Probability Distribution, CPD)。

(1)观测向量节点($\boldsymbol{O}1$ 和 $\boldsymbol{O}2$)：$\boldsymbol{O}1$ 和 $\boldsymbol{O}2$ 分别为音频特征向量和视频特征向量，条件概率 $P(Ox_t \mid Px_t)$, x 为 1 或 2 分别表示音频流和视频流(本文以下部分类同)，采用高斯混合模型来描述。

$$b_{Sx_t}(Ox_t) = f\left(Ox_t \middle| Sx_t\right) = \sum_{k=1}^{M} \omega_{Sx_t k} N\left(Ox_t, \mu_{Sx_t k}, \sigma_{Sx_t k}\right) \quad (1)$$

其中 $\omega_{Sx_t,k}$ 为权值，$\Sigma_k \omega_{Sx_t k} = 1$，$M$ 为混合元个数，$\mu_{Sx_t k}$ 为均值，$\sigma_{Sx_t,k}$ 为协方差。

(2)状态转移概率(ST1 和 ST2), 表示驻留在本状态或转移到下个状态的概率。

(3)状态节点(ST1 和 ST2)：CPD 为 $P(Sx_t \mid \mathrm{SP}x_t, Px_t)$ ，是它的父节点 SPx 和音素 Px 的确定性函数，如果给出了音

素和状态在音素中的位置，那么具体状态就可以得到。表示为

$$p(Sx_t=j\big|Px_t=i,\mathrm{SP}x_t=m)=\begin{cases}1, & j\text{ 为音素}i\quad\text{的第 }m\text{个状态}\\0, & \text{其他}\end{cases}\tag{2}$$

(4)状态在音素中的位置节点(SP1 和 SP2)：在初始帧, $\mathrm{SP}x_1$ 为 0；在其他时间帧，当有音素转移发生时，表示一个音素的结束，状态位置 $\mathrm{SP}x_t$ 值也复位为 0，没有音素转移发生时， $\mathrm{SP}x_t$ 的值由状态转移(STx_t)确定，公式表示为

$$p(\mathrm{SP}x_t=j\,|\,\mathrm{SP}x_{t-1}=i,\mathrm{PT}x_{t-1}=m,\mathrm{ST}x_{t-1}=n)=\begin{cases}1, & m=1,\quad j=0\\1, & m=0,\quad n=1,\quad j=i+1\\1, & m=0,\quad n=0,\quad j=i\\0, & \text{其他}\end{cases}\tag{3}$$

(5)音素节点($P1$ 和 $P2$)：是父节点 PPx 和 W 的确定性函数，该函数确定了词和音素之间的详细关系，对于给定的词，如果给出了音素在词中的位置，那么音素就可以得到。它的 CPD 表示为。

$$p(\mathrm{Px}_t=j\big|W_t=i,\mathrm{PPx}_t=m)=\begin{cases}1, & j\text{是词}i\text{的第}m\text{个音素}\\0, & \text{其他}\end{cases}\tag{4}$$

(6)音素转移概率(PT1 和 PT2)：本文中，每个音素采用了 4 个状态来表示，对于给定的音素(Px)，仅当当前状态为音素的最后一个状态，并且有状态转移发生时，才会有音素转移发生，表示为

$$p(\mathrm{PT}x_t=j\big|Px_t=a,\mathrm{SP}x_t=b,\mathrm{ST}x_t=m)=\begin{cases}1, & j=1,\ m=1,\ b\text{为音素}a\ \text{的最后一个状态}\\1, & j=0,\ m=1,\ b\text{不是音素}a\ \text{的最后一个状态}\\0, & \text{其他}\end{cases}\tag{5}$$

(7)音素在词中的位置节点(PP1 和 PP2)：类似于 SP1 和 SP2，在初始帧, $\mathrm{PP}x_1$ 为 0；在其他帧，当有词转移发生时，表示一个词的结束，$\mathrm{PP}x_t$ 值也复位为 0，没有词转移时，$\mathrm{PP}x_t$ 的值由音素转移概率来确定，公式表示为

$$p(\mathrm{PP}x_t=j\,|\,\mathrm{PP}x_{t-1}=i,\mathrm{WT}_{t-1}=m,\mathrm{PT}x_{t-1}=n)=\begin{cases}1, & m=1,\quad j=0\\1, & m=0,\quad n=1,\quad j=i+1\\1, & m=0,\quad n=0,\quad j=i\\0, & \text{其他}\end{cases}\tag{6}$$

(8)词转移概率节点(WT)：词转移概率由音频流和视频流共同确定，它有 5 个父节点，由于每个词的音素构成不同，需要分别处理，对于给定的词，只有当两个流中的 PPx 都为音素在词中的最后一个音素，而且两流中的音素转移概率 PT1 和 PT2 同时都为 1 时，词转移才会发生。

$$p(\mathrm{WT}_t=j\big|W_t=a,\mathrm{PP1}_t=b,\mathrm{PP2}_t=c,\mathrm{PT1}_t=m,\mathrm{PT2}_t=n)=\begin{cases}1, & j=1,m=1,n=1,b=\mathrm{lastphone1}(a),\\ & c=\mathrm{lastphone2}(a)\\1, & j=0\ \text{and}\ (m\neq1\ \text{or}\ n\neq1\ \text{or}\ b=\sim\mathrm{lastphone1}(a)\\ & \text{or}\ c=\sim\mathrm{lastphone2}(a))\\0, & \text{其他}\end{cases}\tag{7}$$

式中 lastphone1(a) 和 lastphone2(a) 分别表示音频流和视频流中词 a 的最后一个音素。

(9)词节点(W)：在初始帧，词由单文法模型 unigram(i) 确定，而在其他帧，采用了二元文法模型，当没有词转移发生时，词保持不变；当有词转移发生时，由当前词转移到下一个词的概率采用二元文法模型得到。

$$P(W_t=j\,|\,W_{t-1}=i,\mathrm{WT}_t=m)=\begin{cases}\mathrm{bigram}(i,j), & m=1\\1, & m=0,\ i=j\\0, & \text{其他}\end{cases}\tag{8}$$

bigram(i,j) 表示由词 i 转移到 j 的概率，通过对训练样本进行统计得到。

3 识别实验和结果分析

本文采用 GMTK[4]和 HTK 来分别实现本文提到的所有 DBN 模型和 HMM 模型。

3.1 音视频数据库

音视频数据库采用西北工业大学-比利时布鲁塞尔自由大学音视频信号处理联合实验室录制的数字音视频英文数据库和连续音视频英文数据库。数字音视频数据库中有数字 0−10，涉及到 22 个音素(phone)，数据库的脚本按照 Aurora 2.0 语音数据库的句子顺序录制。本文采用 100 句纯净的音视频数据作为训练数据，另外 50 句以及相应加噪语音的音视频数据作为测试数据。对于连续音视频数据库，数据库的脚本采用 TIMIT 数据库生成，本文采用了 600 句音视频数据，包含了 1692 个词和 74 个音素。考虑到样本数据相对比较少，采用 jack-knife 策略，把样本分为两部分，循环进行训练和识别实验，每次采用了 500 句纯净语音的音视频数据进行训练，另外 100 句及加噪语音的音视频数据作为测试样本，最后对 6 次识别结果进行平均。

3.2 音频视频特征提取

音视频特征提取过程见图 3，对音频数据，帧速率为 100 帧/秒，采用 HTK 工具包提取音频数据的 12 维 MFCC 特征和能量特征，加上一阶和二阶差分系数，即 MFCC_D_A，共 42 维音频特征。

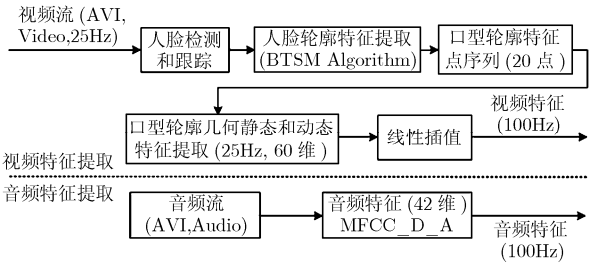

图 3　音频视频特征提取框图

对于视频数据，帧速率为 25 帧/秒，首先进行嘴唇检测和跟踪[7]，然后采用贝叶斯切线形状模型(Bayesian Tangent Shape Model, BTSM)算法[7]进行唇部特征轮廓点的自动标注，基于唇部轮廓特征点，提取唇部的几何特征，包括嘴唇上下左右的张开度(横向和纵向距离)，以及张开时的角度共 20 维特征，最后和第一帧的视频几何特征相减进行归一化处理，同时，为了表示口形动态特征，提取了几何特征的一阶和二阶差分系数，共有 60 维视频特征。最后，为了和音频数据的采样率一致，进行线性插值处理。

3.3 实验安装和结果分析

为评价模型的性能，本文采用了文献[8]中的两个单流 DBN 模型：WP-DBN 和 WPS-DBN 模型(分别为 MS-ADBN 和 MM-ADBN 模型相对应的单流 DBN 模型)。同时采用 HMM 模型和多流异步 HMM(MSHMM, 采用乘积 HMM 实现)在相同的实验条件下进行了语音识别实验。

在连续数字音视频语音识别实验中，对于 WP-DBN 模型[8]，音素和观测向量相联系，采用 1 个高斯模型描述，加上静音和停顿，共有 25 个高斯模型参数需要训练，对于 MS-ADBN 模型，则有 50 个高斯模型参数。而对于 WPS-DBN 模型，每个音素由 4 个状态数构成，状态和观测向量相联系，共有 91 个高斯参数，对于 MM-ADBN 模型，则有 182 个高斯模型参数需要训练。对于训练样本，大约每个词平均有 60 多个训练样本，每个音素平均有 200 个样本，模型可以得到一定的训练，识别结果见表 1。

在连续音视频语音识别实验中，WP-DBN 模型共有 77 个高斯模型参数[8]，MS-ADBN 模型则有 154 个高斯模型参数。而对于 WPS-DBN 模型，共有 299 个高斯参数，对于 MM-ADBN 模型，则有 598 个高斯模型参数需要训练。对于训练样本，大约每个词平均约 3 个训练样本，所以 WP-DBN 模型和 MS-ADBN 模型不能得到充分训练，而每个音素大约有超过 300 个样本，WPS-DBN 模型和 MM-ADBN 模型可以得到一定的训练，识别结果见表 2。

从表 1 和表 2 的结果可以得出下述结论：

(1)由于结合了语音的视觉特征，多流模型的性能明显优于对应的单流模型，对于数字音视频数据库，在信噪比为 0−30dB 的测试环境下，MSHMM, MS-ADBN 模型和 MM-ADBN 模型比对应的单流模型(HMM, WP-DBN 和 WPS-DBN 模型)识别率平均提高 6.03%, 6%和 7%。而对于连续音视频数据库，在纯净语音环境下，识别率分别提高了 1.86%, 2.57%和 5.61%，说明由于视觉特征的辅助作用，提高了系统的识别性能及对噪声的鲁棒性。

(2)对于数字音视频数据库，在信噪比为 0−30dB 的测试环境下，MS-ADBN 模型的识别率比 MSHMM 的识别率平均高 9.93%。对于连续音视频数据库，MM-ADBN 模型的识别率都高于 MSHMM 的识别率，纯净语音下，识别率提高了 9.97%。因为 MS-ADBN 模型和 MM-ADBN 模型在单词之内描述了音频视频流的异步性，而 MSHMM 模型限制音频视频流在音素边界同步。结果表明了音视频异步性的描述对多模态语音识别的重要性。

表 1　数字音视频数据库：实验系统和词识别结果

识 别 系 统	词 识 别 率 (%)							
	0dB	5dB	10dB	15dB	20dB	30dB	Clean	0-30dB
WP-DBN (audio only)	42.94	66.10	71.75	77.97	81.36	96.61	97.74	72.79
HMM (audio only)	30.21	41.0	62.67	74.62	85.67	98.04	98.79	65.36
WPS-DBN (audio only)	19.6	28.7	46.41	64.71	81.7	96.08	97.04	56.2
WP-DBN (video only)	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
HMM (video only)	64.2	64.2	64.2	64.2	64.2	64.2	64.2	64.20
WPS-DBN (video only)	66.06	66.06	66.06	66.06	66.06	66.06	66.06	66.06
MSHMM (audio and visual feature)	44.63	55.31	69.23	77.89	86.92	94.36	95.72	71.39
MS-ADBN (audio and visual feature)	53.94	70.61	86.06	89.39	93.03	95.76	97.27	81.46
MM-ADBN (audio and visual feature)	33.64	43.03	60.61	73.03	81.52	89.39	94.55	63.54

表 2 连续音视频数据库：实验系统和词识别结果

识 别 系 统	词 识 别 率 (%)						
	0dB	5dB	10dB	15dB	20dB	30dB	Clean
WP-DBN (audio only)	2.39	5.61	9.07	14.80	17.06	22.79	27.57
HMM (audio only)	0.72	1.07	3.46	14.32	27.21	44.87	49.76
WPS-DBN (audio only)	2.51	5.13	9.11	16.47	29.24	50.48	62.77
WP-DBN (video only)	6.56	6.56	6.56	6.56	6.56	6.56	6.56
HMM (video only)	10.86	10.86	10.86	10.86	10.86	10.86	10.86
WPS-DBN (video only)	16.11	16.11	16.11	16.11	16.11	16.11	16.11
MSHMM (audio and visual feature)	11.69	18.38	25.89	36.99	44.15	52.15	55.37
MS-ADBN (audio and visual feature)	11.32	12.79	13.18	15.64	17.89	24.10	29.43
MM-ADBN (audio and visual feature)	16.21	21.16	32.72	40.24	49.38	55.98	65.34

(3)对于数字音视频数据库，基于音频视频特征的 WPS-DBN 模型和 MM-ADBN 模型的识别率分别低于 WP-DBN 模型和 MS-ADBN 模型的识别率；相反，对于连续音视频数据库，识别率优于 WP-DBN 和 MS-ADBN 模型，在纯净语音环境下，识别率分别提高了 35.2%和 35.91%。因为在数字音视频数据库下，WP-DBN 和 MS-ADBN 模型可以得到充分的训练，词基元模型优于音素基元模型，而在连续音视频数据库实验中，由于 MM-ADBN 模型和 WPS-DBN 模型的识别基元是音素，可以得到相对充分的训练，而 WP- DBN 模型和 MS-ADBN 模型是整词模型，难以得到充分训练。

(4)对于数字音视频数据库实验，当信噪比小于 20dB 或采用视频特征，WP-DBN 模型识别率都高于 HMM 模型；而对于连续音视频识别实验，WPS-DBN 模型的识别率都高于 HMM 的识别率，虽然 HMM 采用的是三音素模型，而 WPS-DBN 模型采用了单音素的结构，但在纯净语音和视频特征的测试条件下，WPS-DBN 模型的识别率分别提高了 13.01%和 5.52%。可能原因是 DBN 模型能更好描述语音的变化规律，具有更好的识别性能。

(5)虽然多流 DBN 模型性能优于 MSHMM，但是由于 DBN 模型的三角化，推理、搜索的算法还不够完善，特别是应用于连续音视频语音识别的任务时，运行效率不如 MSHMM，距离实用化还需要更深入的研究。

4 结束语

本文提出两个多流动态贝叶斯网络(MS-ADBN 和 MM-ADBN)模型，应用于小词汇量和大词汇量数据库的音视频语音识别，模型放松了音视频流异步性的限制，在词级别上描述了音频视频流的异步性，本质上，MS-ADBN 模型是一个词模型，识别基元是词，而 MM-ADBN 模型是一个音素模型，识别基元是音素。实验结果表明：对于小词汇量的数字音视频数据库。MS-ADBN 模型有最高的识别率，而对于大词汇量连续音视频数据库，纯净语音下，MM-ADBN 模型比 MS-ADBN 模型和多流 HMM 模型的识别率高 35.91%和 9.97%，实验表明了音视频的异步性描述对多模态音视频语音识别系统的重要性。在将来的工作中，我们将继续完善 MM-ADBN 模型，实现三音素捆绑并应用于大词汇量连续音视频数据库的语音识别。

参 考 文 献

[1] Dupont S and Luettin J. Audio-visual speech modeling for continuous speech recognition. *IEEE Trans. on Multimedia*, 2000, 2(3): 141–151.

[2] Potamianos G, and Neti C, *et al*.. Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE*, 2003, 91(9): 1306–1326.

[3] Nefian A, Liang L, and Pi X, *et al*.. Dynamic Bayesian networks for audio-visual speech recognition. EURASIP, *Journal on Applied Signal Processing*, 2002, 2002(11): 1274–1288.

[4] Bilmes J and Zweig G. The graphical models toolkit: An open source software system for speech and time-series processing. In Proc. IEEE Intl. Conf. Acoustics, Speech, and Signal Processing, Orlando, USA, 2002, 4: 3916–3919.

[5] Gowdy J N, Subramanya A, and Bartels C, *et al*.. DBN-based multistream models for audio-visual speech recognition. In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Philadelphia, USA, May 2004, 1: 993–996.

[6] Bilmes J and Bartels C. Graphical model architectures for speech recognition. *IEEE Signal Processing Magazine*, 2005, 22(5): 89–100.

[7] Ravyse Ilse, Jiang D M, and Jiang X Y, *et al*.. DBN based models for audio-visual speech analysis and recognition. 2006 Pacific-Rim Conference on Multimedia (PCM 2006), Hangzhou, China, Nov 2-4, 2006: 19–30.

[8] Lü Guoyun, Jiang Dongmei, and Sahli H, *et al*.. A novel DBN model for large vocabulary continuous speech recognition and phone segmentation. International Conference on Artificial Intelligence and Pattern Recognition (AIPR-07), Orlando, Florida, USA, July 2007: 397–402.

吕国云：　男，1975 年生，博士生，研究方向为模式识别、音视频信号处理.

蒋冬梅：　女，1973 年生，副教授，研究方向为音视频信号处理.

张艳宁：　女，1967 年生，教授，博士生导师，研究方向为视频图像处理和计算机视觉.

赵荣椿：　男，1937 年生，教授，博士生导师，研究方向为语音图像处理和计算机视觉.

H Sahli：　男，教授，研究方向为语音和图像处理.

W Verhelst：　男，教授，研究方向为语音信号处理.