基于多流三音素 DBN 模型的音视频语音识别和音素切分

吕国云[①] 蒋冬梅[①] 樊养余[①] 赵荣椿[①] H. Sahli[②] W. Verhelst[②]
[①](西北工业大学 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：为实现音视频语音识别和同时对音频视频流进行准确的音素切分，该文提出一个新的多流异步三音素动态贝叶斯网络(MM-ADBN-TRI)模型，在词级别上描述了音频视频流的异步性，音频流和视频流都采用了词-三音素-状态-观测向量的层次结构，识别基元是三音素，描述了连续语音中的协同发音现象。实验结果表明：该模型在音视频语音识别和对音频视频流的音素切分方面，以及在确定音视频流的异步关系上，都具备较好的性能。

关键词：语音识别；动态贝叶斯网络；音素切分；音视频
中图分类号：TP391.42 **文献标识码**：A **文章编号**：1009-5896(2009)02-0297-05

DBN Model Based Multi-stream Asynchrony Triphone for Audio-Visual Speech Recognition and Phone Segmentation

Lü Guo-yun[①] Jiang Dong-mei[①] Fan Yang-yu[①]
Zhao Rong-chun[①] H. Sahli[②] W. Verhelst[②]
[①](*Northwestern Polytechnical University*, *Xi'an* 710072, *China*)
[②](*Department ETRO*, *Vrije Universiteit Brussel*, *Brussel*, B-1050, *Belgium*)

Abstract: In this paper, a novel Multi-stream Multi-states Asynchronous Dynamic Bayesian Network based context-dependent TRIphone (MM-ADBN-TRI) model is proposed for audio-visual speech recognition and phone segmentation. The model looses the asynchrony of audio and visual stream to the word level. Both in audio stream and in visual stream, word-triphone-state topology structure is used. Essentially, MM-ADBN-TRI model is a triphone model whose recognition basic units are triphones, which captures the variations in real continuous speech spectra more accurately. Recognition and segmentation experiments are done on continuous digit audio-visual speech database, and results show that: MM-ADBN-TRI model obtains the best overall performance in word accuracy and phone segmentation results with time boundaries, and more reasonable asynchrony between audio and visual speech.

Key words: Speech recognition; Dynamic Bayesian network; Phone segmentation; Audio-visual

1 引言

多模态语音识别和可视语音合成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[1]；构建与语音相一致的说话人面部动画，可以增强人们对语音的理解。然而，虽然人说话时的唇部视觉运动和语音是相关的，但其并不同步，在音视频语音识别模型中应该尽可能体现这种异步性，同时，基于音频视频单元串接的可视语音合成的自然性和逼真性，也在很大程度上取决于对音视频语音数据库中，音频和视频单元及其异步性的正确划分。

对音视频联合建模的研究，文献[1]从各个融合层次进行了分析，文献[1,3]对多流隐马尔可夫模型(HMM)，乘积HMM, Couple HMM, Factorial HMM 等多流 HMM 进行了分析和识别实验。然而目前的多流 HMM 采用音素作为音视频流的建模单元和同步点，仅仅在一定程度上反映音视频流的异步性。而且，所得到的音视频流的音素切分序列是相同的，不能反映音视频流之间音素级的异步关系。

利用动态贝叶斯网络模型(Dynamic Bayesian Network, DBN)进行语音识别已受到人们的关注，与 HMM 相比，DBN 能以图的方式显式地描述语音识别模型，并具有良好的扩展性和可解释性。文献[4,5]将 DBN 模型应用于语音识别。文献[6]给出了一个通用的多流异步 DBN 模型结构，可以在词节点描述音视频流的异步关系，每个流采用了词-状态的层次结构。文献[7]对其进行了改进，音视频流都采用了词-音素的结构，除词识别结果外，还可以得到针对音视频流的音素切分结果。但是本质上仍然为词模型，没有描述音素的动态发音变化过程。

本文以文献[7]中多流 DBN 模型为基本模型，考虑连续语音中的协同发音现象[8]，在音视频流都采用了词-三音素的

2007-07-23 收到，2008-12-11 改回
中国博士后科学基金和中国科技部与比利时弗拉芒大区科技合作项目([2004]487)资助课题

层次结构，构成基于三音素的多流异步 DBN(MS-ADBN-TRI)模型。为更准确地描述三音素的动态发音过程，基于 MS-ADBN-TRI 模型，音视频流都采用了词-三音素-状态的层次结构，构成一个新颖的基于上下文三音素的多流多状态 DBN(MM-ADBN-TRI)模型。最后进行音视频语音识别和音素切分实验，并分析音视频流之间的异步关系。

2 多流异步三音素 DBN 模型

基于文献[7]中的模型，考虑连续语音中的协同发音现象[8]，不改变模型的整体结构，仅仅把音频流和视频流中的词-音素-观测向量的组成结构改变为词-三音素-观测向量的组成结构，相应的音素节点、音素转移概率节点更改为三音素节点、三音素转移节点，构成 MS-ADBN-TRI 模型，更精确地描述了词的动态发音变化过程，但没有反映三音素的动态发音过程。MM-ADBN-TRI 模型是 MS-ADBN-TRI 模型的扩展，在音视频流的观测向量和三音素之间增加了状态节点层，每个三音素由固定个数的状态构成，状态和观测向量联系，它的识别模型见图 1 所示。可以看到，词和词转移变量位于模型的上方，被音视频流所共享，音视频流在词节点同步，而在两个词之间，音视频流都采用了词-三音素-状态-观测向量的层次结构，识别基元是三音素。圆括号内为对应节点变量的简称。

图 2 中的(a)和(b)分别描述了 MS-ADBN-TRI 模型和 MM-ADBN-TRI 模型的层次结构(并给出词 five 的实例)。可以看到，前者是词模型，没有反映三音素的动态变化过程；而后者识别基元是三音素，不但描述了词的动态变化过程，还描述了三音素的动态变化过程。

为更好的理解模型，下面描述 MM-ADBN-TRI 模型的主要节点变量及其条件概率分布。

(1)音频和视频观测向量($\boldsymbol{O}1$ 和 $\boldsymbol{O}2$)。条件概率为 $P(Ox_t \mid Sx_t)$，x 为 1 或 2 分别表示音频流和视频流(本文以下部分

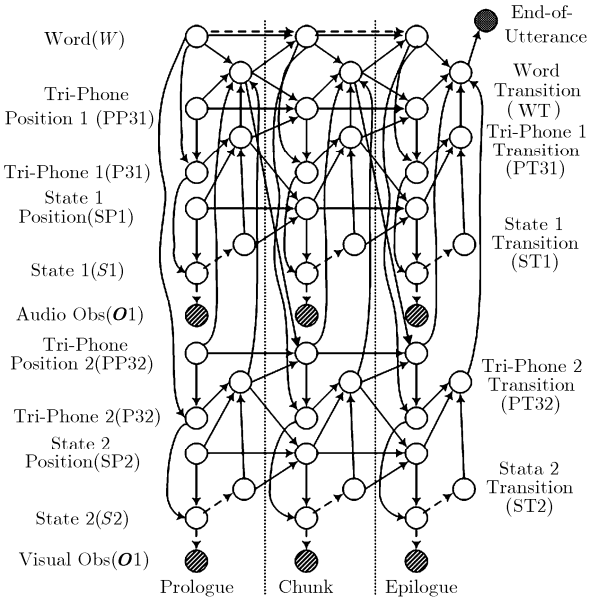

图 1 MM-ADBN-TRI 连续语音识别和音素切分模型

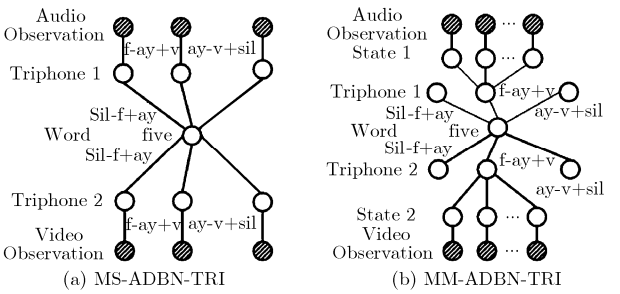

图 2 多流 DBN 模型的节点层次关系示意图

类同)

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

其中 $\omega_{Sx_t,k}$ 为权值，$\Sigma_k \omega_{Sx_t k} = 1$，$\mu_{Sx_t,k}$ 为均值，$\sigma_{Sx_t,k}$ 为协方差。M 为高斯混合模型的混合元个数。

(2)状态转移(STx)，状态节点(Sx)和状态在三音素中的位置(SPx)节点

$$P(\mathrm{ST}x_t = b \mid Sx_t = i) = \begin{cases} A_{ii}, & b = 0 \\ 1 - A_{ii}, & b = 1 \end{cases} \quad (2)$$

$$p(Sx_t = j \middle| P3x_t = i, \mathrm{SP}x_t = m) = \begin{cases} 1, & j\text{为三音素}i\text{的第}m\text{个状态} \\ 0, & \text{其他} \end{cases} \quad (3)$$

$$p(\mathrm{SP}x_t = j \mid \mathrm{SP}x_{t-1} = i, \mathrm{PT3}x_{t-1} = m, \mathrm{ST}x_{t-1} = n) = \begin{cases} 1, & m = 1, j = 0, t \neq 0 \\ 1, & m = 0, n = 1, j = i+1, t \neq 0 \\ 1, & m = 0, n = 0, j = i, t \neq 0 \\ 1, & j = 0, t = 0 \\ 0, & \text{其他} \end{cases} \quad (4)$$

其中 A_{ii} 表示呆在状态 i 的概率，而 $1 - A_{ii}$ 表示从状态 i 转移到状态 i+1 的概率。

(3)三音素($P3x$)，三音素转移(PT3x)和三音素在词中的位置(PP3x)节点

$$p(P3x_t = j \middle| W_t = i, \mathrm{PP3}x_t = m) = \begin{cases} 1, & j\text{是词}i\text{的第}m\text{个三音素} \\ 0, & \text{其他} \end{cases} \quad (5)$$

$$p(\mathrm{PT3}x_t = j \middle| P3x_t = a, \mathrm{SP}x_t = b, \mathrm{ST}x_t = m) = \begin{cases} 1, & j{=}1, m{=}1, b\text{为三音素}a\text{的最后一个状态} \\ 1, & j{=}0, m{=}1, b\text{不是三音素}a\text{的最后一个状态} \\ 0, & \text{其他} \end{cases} \quad (6)$$

$$p(\mathrm{PP3}x_t = j \mid \mathrm{PP3}x_{t-1} = i, \mathrm{WT}_{t-1} = m, \mathrm{PT3}x_{t-1} = n) = \begin{cases} 1, & m = 1, j = 0, t \neq 0 \\ 1, & m = 0, n = 1, j = i+1, t \neq 0 \\ 1, & m = 0, n = 0, j = i, t \neq 0 \\ 1, & j = 0, t = 0 \\ 0, & \text{其他} \end{cases} \quad (7)$$

(4)词转移节点(WT)和词节点(W)：对给定的词，当音视频流中的 PP3x 都为当前词中的最后一个三音素，且两流中的三音素转移 PT31 和 PT32 同时发生时，词转移才发生(WT=1)，公式表示为

$$\begin{aligned}&p(\mathrm{WT}_t=j\big|W_t=a,\mathrm{PP31}_t=b,\mathrm{PP32}_t=c,\mathrm{PT31}_t\\&\quad=m,\mathrm{PT32}_t=n)\\&=\begin{cases}1, & j=1,m=1,n=1,b=\text{lasttriphone1}(a),\\ & c=\text{lasttriphone2}(a)\\ 1, & j=0 \text{ and } (m\neq 1 \text{ or } n\neq 1 \text{ or } b=\\ & \sim \text{lasttriphone1}(a) \text{ or } c=\sim \text{lasttriphone2}(a))\\ 0, & \text{其他}\end{cases}\end{aligned}\tag{8}$$

式中 lasttriphone1(a) 表示音频流中当前三音素是词 a 最后一个三音素，lasttriphone2(a) 表示视频流中当前三音素是词 a 最后一个三音素。另外，词节点和文献[7]中相同，当有词转移发生时，当前词转移到下一个词的概率采用二元文法模型。

3　识别和音素切分实验

3.1 音视频数据库与特征提取

音视频数据库采用中国-比利时听视觉信号处理联合实验室录制的连续数字音视频英文数据库，数据库在安静的环境下录制，视频图像为正面头部视频图像，数据库中有数字 0–10，涉及到 22 个单音素(词和音素的组成采用 TIMIT 数据库的表示方式)和 36 个三音素单元(词内上下文扩展得到)，数据库的脚本按照 Aurora 2.0 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 句以及相应的加噪语音的音视频数据作为测试数据。

对于音频数据，帧速率为 100 帧/秒，提取 13 维 Mel 倒谱系数和能量参数，并结合其一阶和二阶差分系数，形成 42 维语音特征。

对于视频数据，帧速率为 25 帧/秒，首先进行唇部检测和跟踪[4]，然后采用贝叶斯切线形状模型算法[9]进行唇部轮廓点的自动标注，基于唇部轮廓点，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，并和每句话的第一帧几何特征相减进行归一化；同时，提取这些特征的一阶和二阶差分系数，形成 60 维视频特征[7]。最后，为了和音频数据的帧速率一致，进行了线性插值处理。

3.2 实验设置和结果分析

为了测试 MM-ADBN-TRI 和 MS-ADBN-TRI 模型的语音识别和音素切分性能，与 MM-ADBN-TRI 和 MS-ADBN-TRI 模型相对应的单流 DBN 三音素模型(WPS-DBN-TRI, WP-DBN-TRI 为文献[5]中 WP-DBN 和 WPS-DBN 的改进，词-音素的构成方式更改为词-三音素的构成方式)被采用来进行比较。建模时，WPS-DBN-TRI 和 MM-ADBN-TRI 模型，共有 36 个三音素单元，每个三音素采用 4 个状态来描述。同时，采用常规 HMM 和文献[1]中提到的多流 HMM (Multi-stream HMM, MSHMM)，在相同的实验环境下也进行了识别实验。MSHMM 模型采用乘积 HMM(Product HMM, PHMM)实现。模型训练中，每个词大约平均有 60 多个训练样本，而每个音素平均有 200 个训练样本，每个三音素大约平均有 150 多个样本，所有模型都可以得到适当的训练。而本文采用 GMTK[6]工具包对所有 DBN 模型进行训练，并得到语音识别和带时间边界的音素切分结果。

3.2.1 语音识别实验　MS-ADBN-TRI 和 MM-ADBN-TRI 模型的词识别结果见表 1，作为比较，相同实验环境下，HMM, WP-DBN-TRI, WPS-DBN-TRI 模型和 MSHMM 的词识别率也在表 1 中给出，可以得到下面的结论：

(1)由于结合了说话时的唇部视觉特征，多流模型的性能高于对应的单流模型，体现了更好的噪声鲁棒性。在信噪比为 0–30dB 的测试环境下，MS-ADBN-TRI 模型和 MM-ADBN-TRI 模型的识别率分别比对应的 WP-DBN-TRI 模型和 WPS-DBN-TRI 模型的识别率平均高 6.72%和 5.84%。

(2)在 0–30dB 的语音测试环境下，MS-ADBN-TRI 和 MM-ADBN-TRI 模型的识别率都高于 MSHMM 模型的识别率。这是因为 MSHMM 模型限制音频视频流必须在音素级同步，而 MS-ADBN-TRI 和 MM-ADBN-TRI 模型放松了音频视频的异步性限制，描述了单词内音视频的异步性，结果

表 1 词识别率：DBN 模型和 HMM 模型的比较

系　统	识　别　率 (%)							
	0dB	5dB	10dB	15dB	20dB	30dB	Clean	0–30dB
WP-DBN-TRI　(audio only)	45.38	68.32	73.11	82.32	83.79	98.36	99.17	75.21
WPS-DBN-TRI　(audio only)	32.31	46.25	64.37	74.98	82.13	97.73	98.60	66.29
WP-DBN-TRI　(video only)	67.86	67.86	67.86	67.86	67.86	67.86	67.86	67.86
WPS-DBN-TRI　(video only)	66.72	66.72	66.72	66.72	66.72	66.72	66.72	66.72
MSHMM　(audio and visual feature)	44.63	55.31	69.23	77.89	86.92	94.36	95.72	71.39
MS-ADBN-TRI　(audio and visual feature)	54.34	70.68	86.41	90.27	93.13	96.76	97.35	81.93
MM-ADBN-TRI　(audio and visual feature)	48.34	57.38	69.50	78.12	85.52	93.91	95.57	72.13

也表明音视频异步性描述对于多模态语音识别结果的影响。

(3)在 0–30dB 的语音测试环境和基于视频特征的测试环境下，WP-DBN-TRI 模型和 MS-ADBN-TRI 模型分别比 WPS-DBN-TRI 和 MM-ADBN-TRI 模型有更高的语音识别率，一个可能的原因是：前者都是词模型，而后者是三音素模型，对于小词汇量语音识别，在词基元可以得到较好训练的情况下，识别基元为词的模型性能优于识别基元为三音素的模型。

3.2.2 音频视频流音素切分实验 为了评价针对音频流的音素切分结果，采用文献[5]中提到的一个客观的音素切分评价标准：音素切分正确性(Phone Segmentation Accuracy, PSA)。DBN 模型得到的音素切分结果和参考序列逐帧比较，PSA=A/C，A 为相同的帧数，C 为总帧数。而为了评价针对视频流的音素切分结果，本文提出一个客观评价标准：视素切分正确性(Viseme Segmentation Accuracy, VSA)的。考虑到音频视频流的异步性，为给出一个准确的视素参考序列，对测试集的视频流进行手工切分，在切分时，同时观察每句话音频流的音素变化和视频流的口形变化，先得到针对视频流的参考的音素切分序列，然后按照文献[10]中的音素-视素对应关系得到参考的视素序列。由于 DBN 模型得到的是针对视频流的音素切分序列，也采用文献[10]中音素-视素的影射关系得到视素切分序列，对于所得到的视素序列，逐帧和视素参考序列进行比较，得到 VSA=B/C，B 为正确的帧数，C 为总帧数。

对于所有的测试集，可以得到平均的 PSA 和 VSA。结果见表 2，可以得到下面的结论。

表 2 音视频流：不同信噪比下的平均 PSA 和 VSA (%)

系 统	平均 PSA						平均 VSA					
	5dB	10dB	15dB	20dB	30dB	Clean	5dB	10dB	15dB	20dB	30dB	Clean
WP-DBN-TRI (audio or video)	42.8	45.8	55.3	64.1	82.4	86.8	52.3	52.3	52.3	52.3	52.3	52.3
WPS-DBN-TRI(audio or video)	41.7	44.3	59.5	67.1	74.9	89.8	57.1	57.1	57.1	57.1	57.1	57.1
MS-ADBN-TRI (audio and video)	28.1	30.4	35.3	36.2	40.4	43.5	54.8	57.3	60.5	62.8	64.6	69.2
MM-ADBN-TRI (audio and video)	36.2	42.3	49.9	57.2	64.9	81.7	51.8	56.2	60.9	63.1	66.3	75.6

(1) WP-DBN-TRI 和 WPS-DBN-TRI 模型的平均 PSA 值分别高于对应的 MS-ADBN-TRI 和 MM-ADBN-TRI 模型的平均 PSA。这是因为音素参考序列是从纯音频数据得到的，而多流 DBN 模型(MS-ADBN-TRI 和 MM-ADBN-TRI 模型)强制音频流和视频流在词的边界上同步，切分出的音素时间边界有所改变，而单流 DBN 模型却没有这样的限制。

(2)当信噪比大于 10dB 时，MS-ADBN-TRI 和 MM-ADBN-TRI 模型的平均 VSA 值分别高于对应 WP-DBN-TRI 和 WPS-ADBN-TRI 模型，原因很明显，对于多流 DBN 模型，在高信噪比的环境下，由于音频特征流的影响，提高了对视频流的视素切分正确性。

(3)又一个有趣的结果是 MM-ADBN-TRI 模型在各信噪比下的平均 PSA 和平均 VSA 值都明显高于 MS-ADBN-TRI 模型，而 MS-ADBN-TRI 模型的音素切分结果几乎是不能被接受的。一个最主要的原因是 MM-ADBN-TRI 模型本质上是一个三音素模型，描述了音素的动态发音变化过程，而 MS-ADBN-TRI 模型是一个词模型，对三音素的描述仅仅是一个静态描述。

3.2.3 音视频流异步关系的定性分析 根据 3.3.2 部分针对音视频流的音素切分结果，可以得到基于音素的音频流和视频流的异步关系。然而，这种异步关系与音素识别率、以及对音频流和视频流的音素切分结果相关，难以找到一个客观的评价标准。作为一个特例，在图 3 中，列出在纯净语音环境下，对一句话“Sil-two-nine-Sil”的音视频流的音素切分结果，定性分析音视频流的异步关系，可得出下面的结论：

(1)从参考的音视频流的音素序列可知：(a)语音的音频流和视频流是不同步的，视觉运动总先于声音信号，并且异步性是非常复杂的，它随着不同音素、不同时间的改变而改变。(b)一些音素的音频视频的异步关系已经超越了音素的边界(比如音素't'和第一个音素'n')。

(2)WP-DBN-TRI 和 WPS-DBN-TRI 模型对音视频流的音素切分是不受限制的，音视频流之间的异步性很大，一些音素的音频视频之间的异步是不合理的(音素'ay1'以及第一个'n')；而 MS-ADBN-TRI 与 MM-ADBN-TRI 模型所得

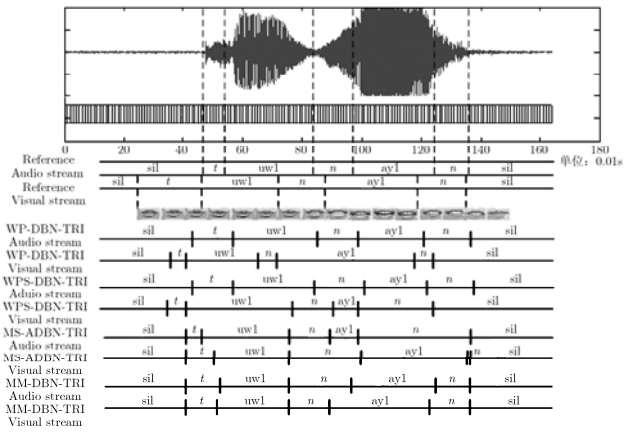

图 3 基于 DBN 模型的音视频流异步性比较：对同一句话“Sil-two-nine-Sil”

到音频视频流之间的异步关系，由于受到模型结构的约束而在词的边界同步，使得音视频流之间的切分结果有相互吸引而靠近的趋势。

(3)MM-ADBN-TRI 模型得到了比 MS-ADBN-TRI 模型得到了更合理的音视频流的异步性。而从 MS-ADBN-TRI 的音素切分结果来看，对一些音素单元，它们的音频竟然先于视频，如第 2 个音素'n'的视频落后于音频大约 370ms，这和唇部运动经常先于声音的结论显然是矛盾的。

上述结论仅仅是对一个范例的定性分析。MS-ADBN-TRI 模型在纯净语音下对音频流的音素切分结果仅仅 43.5%，WP-DBN-TRI 和 WPS-DBN-TRI 模型对基于视频特征流的识别和切分结果都不理想。而 MM-ADBN-TRI 模型在词识别率、对音视频流的音素切分结果方面都具有较好的性能，并能较好地描述音视频流的异步关系，是一个值得深入研究的模型。

另外本文所提出的多流 DBN 模型，特别是 MM-ADBN-TRI 模型，虽然可以得到很好的性能，但是模型训练和识别花费时间大约为 MSHMM 的十倍以上，因此，为了在实际系统中得到应用，还需要在 DBN 的三角化，快速推理和搜索算法等方面进行深入的研究。

4　结束语

基于文献[7]中的多流 DBN 模型，考虑连续语音中的协同发音现象，本文提出了两个多流三音素 DBN 模型：MS-ADBN-TRI 模型和 MM-ADBN-TRI 模型。但是两者组成结构不同，前者属于一个词模型，音视频流都采用词-三音素的层次结构；后者是一个三音素模型，音视频流都采用了词-三音素-状态的层次结构，识别基元为三音素，描述了三音素的动态发音变化过程。最后采用连续数字音视频英文数据库进行了识别和音素切分实验，结果表明：MS-ADBN-TRI 和 MM-ADBN-TRI 模型描述了单词内音频视频流的异步性，识别率都高于 MSHMM(描述了音素内音视频的异步性)，而 MM-ADBN-TRI 模型在词识别率和对音视频流的音素切分方面具备最佳的整体性能，并且可以得到音视频流之间合理的异步关系。另外 MM-ADBN-TRI 模型是一个三音素模型，可以用于对大词汇量的音视频语音识别和音素切分。在将来的工作中，我们将应用 MM-ADBN-TRI 模型进行大词汇量音视频数据库的语音识别研究，而且同时对音视频语音数据库进行音素切分，得到音素单元下音频视频片段数据库，为进行逼真的可视语音合成奠定基础。

参 考 文 献

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

[2] 王志明, 蔡莲红, 艾海舟. 基于数据驱动方法的汉语文本-可视语音合成. 软件学报, 2005, 16(6): 1054–1063.
Wang Z M, Cai L H, and Ai H Z. Text-to-visual speech in Chinese based on data-driven approach. *Journal of Software*, 2005, 16(6): 1054–1063.

[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] 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, 2006: 19–30.

[5] 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.

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

[7] Lü Guoyun, Jiang Dongmei, and Zhao Rongchun, *et al.*. Multi-stream asynchrony Dynamic Bayesian Network model for audio-visual continuous speech recognition. 14th International Conference on systems, Signals and Image Processing (IWSSIP 2007), Maribor, Slovenia, June, 2007, 1: 437–440.

[8] Young S J, Odell J, and Woodland P C. Tree-based state tying for high accuracy acoustic modeling. In Proceedings ARPA Workshop on Human Language Technology, Plainsboro, New Jersey, USA, 1994: 307–312.

[9] Zhou Yi, Gu Lie, and Zhang Hongjiang. Bayesian tangent shape model: Estimating shape and pose parameters via bayesian inference. The IEEE Conference on Computer Vision and Pattern Recognition, Wisconsin, USA, June, 2003, 1: 109–116.

[10] Jiang D M, Xie L, and Zhao R C, *et al.*. Acoustic viseme modeling for speech driven animation: a case study. In Proc. 1st IEEE Benelux Workshop on Model based Processing and coding of Audio (MPCA-2002), Leuven, Belgium, November, 2002, 1: 49–52.

吕国云：　男，1975 年生，博士后，研究方向为音视频信号处理、模式识别、虚拟现实.

蒋冬梅：　女，1974 年生，副教授，研究方向为音视频信号处理.

樊养余：　男，1960 年生，教授，博士生导师，研究方向为视频图像处理、虚拟现实.

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，1960 年生，教授，研究方向为语音和数字图像处理.

W.Verhelst：　男，1961 年生，教授，研究方向为数字语音信号处理.