Audio Visual Speech Recognition and Segmentation Based on DBN Models

Dongmei Jiang[1,2], Guoyun Lv[1], Ilse Ravyse[2], Xiaoyue Jiang[1],
Yanning Zhang[1], Hichem Sahli[2,3] and Rongchun Zhao[1]
Joint NPU-VUB Research Group on Audio Visual Signal Processing (AVSP)
[1] Northwestern Polytechnical University (NPU), School of computer Science,
127 Youyi Xilu, Xi'an 710072, P.R.China
[2] Vrije Universiteit Brussel (VUB), Electronics & Informatics Dept.,
VUB-ETRO, Pleinlaan 2, 1050 Brussels, Belgium
[3] IMEC, Kapeldreef 75, 3001 Leuven, Belgium

1. Introduction

Automatic speech recognition is of great importance in human-machine interfaces. Despite extensive effort over decades, acoustic-based recognition systems remain too inaccurate for the vast majority of real applications, especially those in noisy environments, e.g. crowed environment. The use of visual features in audio-visual speech recognition is motivated by the speech formation mechanism and the natural speech ability of humans to reduce audio ambiguities using visual cues. Moreover, the visual information provides complementary cues that cannot be corrupted by the acoustic noise of the environment. However, problems such as the selection of the optimal set of visual features, and the optimal models for audio-visual integration remain challenging research topics. In recent years, the most common model fusion methods for audio visual speech recognition are Multi-stream Hidden Markov Models (MSHMMs) such as product HMM and coupled HMM. In these models, audio and visual features are imported to two or more parallel HMMs with different topology structures. These MSHMMs describe the correlation of audio and visual speech to some extent, and allow asynchrony within speech units. Compared with the single stream HMM, system performance is improved especially in noisy speech environment. But at the same time, problems remain due to the inherent limitation of the HMM structure, that is, on some nodes, such as phones, syllables or words, constraints are imposed to limit the asynchrony between audio stream and visual stream to phone (or syllable, word) level. Since for large vocabulary continuous speech recognition task, phones are the basic modeling units, audio stream and visual stream are forced to be synchronized at the timing boundaries of phones, which is not coherent with the fact that the visual activity often precedes the audio signal even by 120 ms.

Besides the audio visual speech recognition to improve the word recognition rate in noisy environments, the task of audio visual speech units (such as phones or visemes) segmentation also requires a more reasonable speech model which describes the inherent correlation and asynchrony of audio and visual speech.

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

Dynamic Bayesian Network (DBN) is a good speech model due to its strong description ability and flexible structure. DBN is a statistic model that can represent multiple collections of random variables as they evolve over time. Coupled HMM and product HMM are special cases of much more general DBN models. In fact, many DBN models have been proposed in recent years for speech recognition. For example, in [Zweig 1998], a baseline DBN model was designed to implement a standard HMM, while in [Bilmes 2001], a single stream DBN model with whole-word-state structure, i.e a fixed number of states are assigned to one word to assemble the HMM of the word, was designed for small vocabulary speech recognition. Experimental results show that it obtains high word recognition performance and robustness to noise. But the DBN model with whole-word-state structure, does not allow to make a speech subunit segmentation. [Zhang 2003] extended the single stream whole-word-state DBN model to multi-stream inputs with different audio features such as Mel filterbank cepstrum coefficients (MFCC) and perceptual linear prediction (PLP) features, and built two multi-stream DBN (MSDBN) models: i) a synchronous MSDBN model which assumes that multiple streams are strictly synchronized at the lowest state level, namely, all the different types of feature vectors on the same time frame are tied to one state variable; ii) an asynchronous MSDBN model which allows for limited asynchrony between streams at the state level, forcing the two streams to be synchronized at the word level by resetting the state variables of both streams to their initial value when a word transition occurs. [Gowdy 2004] combined the merit of the synchronous and asynchronous MSDBN models on a mixed type MSDBN model, and extended the speech recognition experiments on an audio visual digit speech database. In the mixed type MSDBN model, on each time slice, all the audio features share one state variable, while the visual stream and the composite audio stream each depend on different state variables which introduces the asynchrony between them. [Bilmes 2005] introduced a new asynchronous MSDBN model structure in which the word transition probability is determined by the state transitions and the state positions both in the audio stream and in the visual stream. Actually this structure assumes the same relationship between word transition and states with the asynchronous MSDBN model in [Zhang 2003], but describes it by different conditional probability distributions. Despite their novelties in breaking through the limitation of MSHMMs by allowing the asynchrony between streams to exceed the timing boundaries of states in a word, all the MSDBN models above build the whole-word-state model for a word i.e emulate the HMM that a word is composed of a fixed number of states, and therefore, the relationships between words and their corresponding subword units (for example phones) are not described. As a consequence, no subword unit level recognition and segmentation output can be obtained from these synchronous and asynchronous MSDBN models.

[Bilmes 2005] also presented a single stream bigram DBN model, in which the composing relationship between words and their corresponding phone units are defined by conditional probability distributions. This model emulates another type of word HMM in which a word is composed of its corresponding phone units instead of fixed number of states, by associating each phone unit with the observation feature vector, one set of Gaussian Mixture Model (GMM) parameters are trained. This model gives the opportunity to output the phone units with their timing boundaries, but to our best knowledge, no experiments have been done yet on evaluating its recognition and segmentation performance of the phone units (or viseme units in visual speech). Further more, it has not been extended to a synchronous or asynchronous multi-stream DBN model to emulate the word composition of its subword units simultaneously in audio speech and visual speech.

In our work, the new contributions to the speech modeling are: 1) the single stream DBN (SDBN) model in [Bilmes 2005] is implemented, speech recognition and segmentation experiments are done on audio speech and visual speech respectively. Besides the word recognition results, phone recognition and segmentation outputs are also obtained for both audio and visual speech. 2) a novel multi-stream asynchronous DBN (MSADBN) model is designed, in which the composition of a word with its phone units in audio stream and visual stream is explicitly described by the phone transitions and phone positions in each stream, as well as the word transition probabilities decided by both streams. In this MSADBN model, the asynchrony of audio speech and visual speech exceeds the timing boundaries of phone units but is restricted to word level. 3) for evaluating the performance of the single stream and multi-stream asynchronous DBN models, besides the word recognition rate, recognition and segmentation accuracy of the phone units with their timing boundaries in the audio stream is compared to the results from the well trained triphone HMMs, segmentation of visemes in the visual stream is compared to the manually labeled references, and the asynchrony between the segmented phone and viseme units is also analyzed.

The sections are organized as follows. Section 2.1 discusses the visual features extraction, starting from the detection and tracking of the speaker's head in the image sequence, followed by the detailed extraction of mouth motion, and section 2.2 lists the audio features. The structures of the single stream DBN model and the designed multi-stream asynchronous DBN model, as well as the definitions of the conditional probability distributions, are addressed in section 3. While section 4 analyzes the speech recognition and phone segmentation results in the audio and visual stream obtained by the SDBN and the MSADBN model, concluding remarks and future plans are outlined in section 5.

2. Audio-Visual Features Extraction

2.1 Visual Feature Extraction

Robust location of the speaker's face and the facial features, specifically the mouth region, and the extraction of a discriminant set of visual observation vectors are key elements in an audio-video speech recognition system. The cascade algorithm for visual feature extraction used in our system consists of the following steps: face detection and tracking, mouth region detection and lip contour extraction for 2D feature estimation. In the following we describe in details each of these steps.

Head Detection and Tracking The first step of the analysis is the detection and tracking of the speaker's face in the video stream. For this purpose we use a previously developed head detection and tracking method [Ravyse 2006]. The head detection consists of a two-step process: (a) face candidates selection, carried out here by iteratively clustering the pixel values in the YC_rC_b color space and producing labeled skin-colored regions $\{R_i\}_{i=1}^{N}$ and their best fit ellipse $E_i = (x_i, y_i | a_i, b_i, \theta)$ being the center coordinates, the major and minor axes length, and the orientation respectively, and (b) the face verification that selects the best face candidate. In the verification step a global face cue measure M_i, combining gray-tone cues and ellipse shape cues, is estimated for each face candidate region R_i. Combining shape and facial feature cues ensures an adequate detection of the face. The face candidate that has the maximal measure M_i localizes the head region in the image.

The tracking of the detected head in the subsequent image frames is performed via a kernel-based method wherein a joint spatial-color probability density characterizes the head region [Ravyse 2005].

Fig. 1 illustrates the tracking method. Samples are taken from the initial ellipse region in the first image, called *model target*, to evaluate the model target joint spatial-color kernel-based probability density function (p.d.f.). A hypothesis is made that the true target will be represented as a transformation of this model target by using a motion and illumination change model. The hypothesized target is in fact the modeled new look in the current image frame of the initially detected object. A *hypothesized target* is therefore represented by the *hypothesized p.d.f.* which is the transformed model p.d.f. To verify this hypothesis, samples of the next image are taken within the transformed model target boundary to create the *candidate target* and the joint spatial-color distribution of these samples is compared to the *hypothesized p.d.f.* using a distance-measure. A new set of transformation parameters is selected by minimizing the distance-measure. The parameter estimation or tracking algorithm lets the target's region converge to the true object's region via changes in the parameter set.

This kernel-based approach is proved to be robust, and moreover, incorporating an illumination model into the tracking equations enables us to cope with potentially distracting illumination changes.

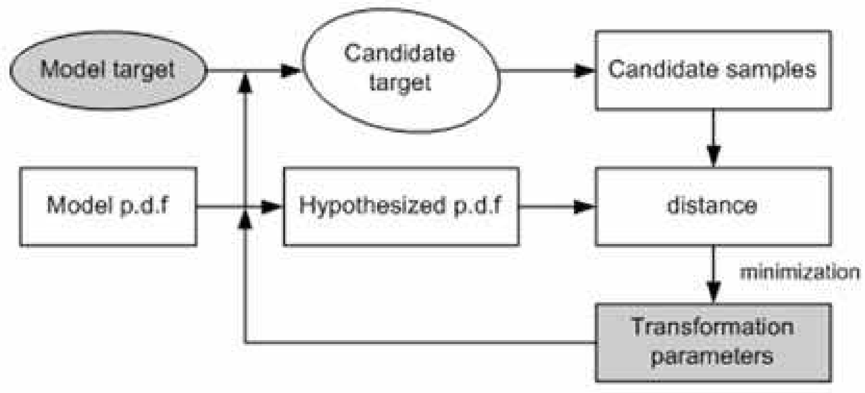


Figure 1. Tracking algorithm

2D Lip Contour Extraction The contour of the lips is obtained through the Bayesian Tangent Shape Model (BTSM) [Zhou 2003]. Fig. 2 shows several successful results of the lip contour extraction.

The lip contour is used to estimate a visual feature vector consisting of the mouth opening measures shown in Fig. 3. In total, 42 mouth features have been identified based on the automatically labeled landmark feature points: 5 vertical distances between the outer contour feature points; 1 horizontal distance between the outer lip corners; 4 angles; 3 vertical distances between the inner contour feature points; 1 horizontal distance between the inner lip corners; and the first order and second order regression coefficient (delta and acceleration in the image frames at 25 fps) of the previous measures.

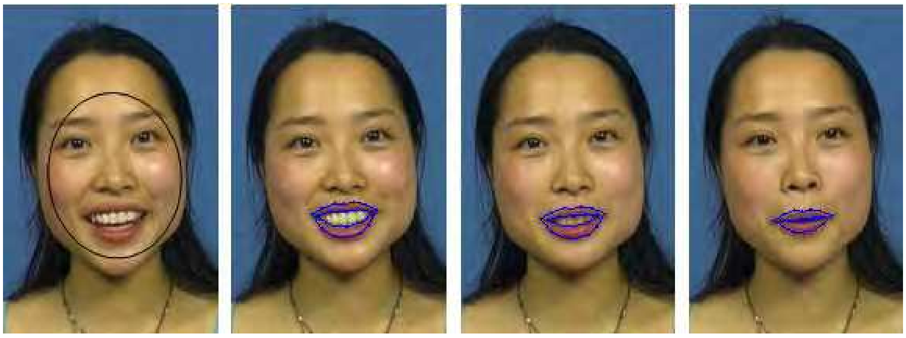
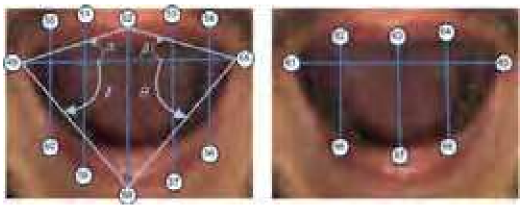


Figure 2. Face detection/tracking and lip contour extraction



(a) (b)

Figure 3. Vertical and horizontal opening distances and angle features of the mouth: (a) outer contour features; (b) inner contour features

2.2 Audio Feature Extraction

The acoustic features are computed with a frame rate of 100frames/s. In our experiments, two different types of acoustic features are extracted: (i) 39 MFCC features: 12 MFCCs [Steven 1980], energy, together with their differential and acceleration coefficients; (ii) 39 dimension PLP features: 12 PLP coefficients [Hermansky 1990], energy, together with their differential and acceleration coefficients.

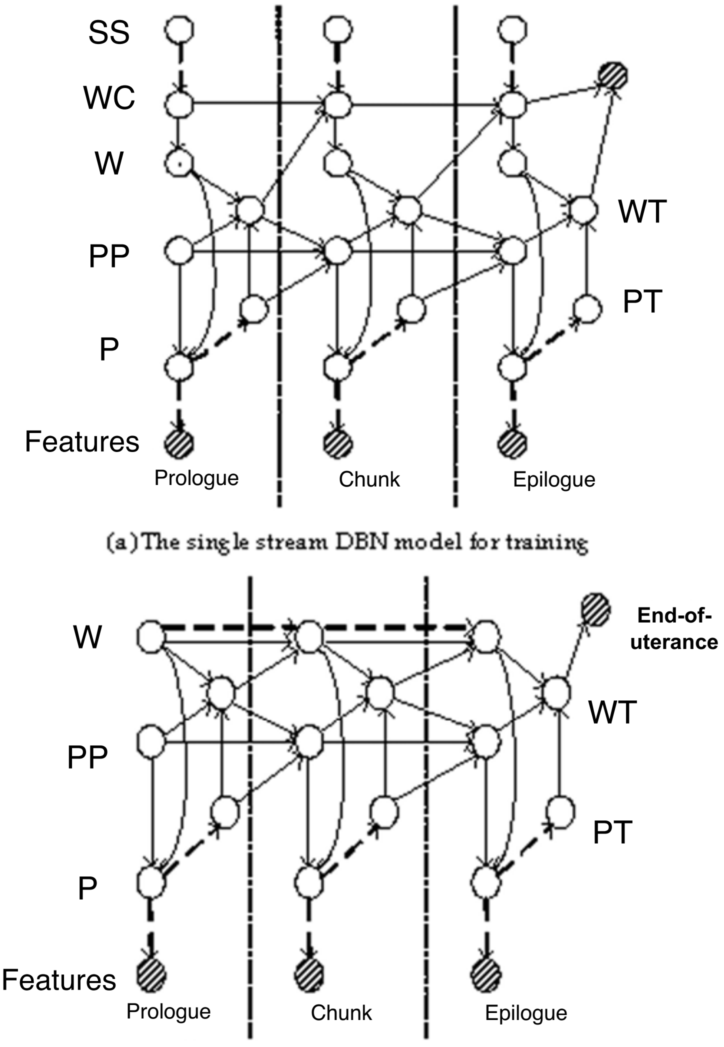
3. Single Stream DBN Model and Multi-Stream Asynchronous DBN Model

3.1 Single Stream DBN Model (SDBN)

In our framework, we first implement the single-stream DBN model following the idea of the bigram DBN model in [Bilmes 2005], and adopt it to the segmentation of phone units and viseme units both in the audio speech and in the visual speech respectively. The training data consists of the audio and video features extracted from word labeled speech sentences.

The DBN models in Fig. 4 represent the unflattened and hierarchical structures for a speech recognition system. (a) is the training model and (b) the decoding model. They consist of an initialization with a *Prologue* part, a *Chunk* part that is repeated every time frame (*t*), and a closure of a sentence with an *Epilogue* part. Every horizontal row of nodes in Fig. 4 depicts a separate temporal layer of random variables. The arcs between the nodes are either deterministic (straight lines) or random (dotted lines) relationships between the random variables, expressed as conditional probability distributions (CPD).

In the training model, the random variables *Word Counter* (WC) and *Skip Silence* (SS), denote the position of the current word or silence in the sentence, respectively. The other random variables in Fig. 4 are: (I) the word identity (W); (II) the occurrence of a *transition to another word* (WT), with *WT* = 1 denoting the start of a new word, and *WT* = 0 denoting the continuation of the current word; (III) the position of the current phone in the current word (PP); (iv) the occurrence of a *transition to another phone* (PT), defined similarly as WT; and (v) the *phone identification* (P), e.g. 'f' is the first phone in the word 'four'.



(a) The single stream DBN model for training

(b) The single stream DBN model for decoding

Figure 4. The single stream DBN model

Suppose the input speech contains T frames of features, for the decoding model of SDBN as shown in Fig.4, the set of the all the hidden nodes is denoted as $H_{1:T} = \left(W_{1:T}, WT_{1:T}, PP_{1:T}, PT_{1:T}, P_{1:T}\right)$, then the probability of observations can be computed as

$$P\left(O_{1:T}\right) = \sum_{H_{1:T}} P\left(H_{1:T}, O_{1:T}\right)$$

The graph thus specifies the following factorization for the joint probability distribution as:

$$\begin{aligned} &P(H_{1:T}, O_{1:T}) = P(W_{1:T}, WT_{1:T}, PP_{1:T}, PT_{1:T}, P_{1:T}, O_{1:T}) = \\ &\prod_{t=1}^{T} P(O_t|P_t) \cdot P(P_t|PP_t, W_t) \cdot P(PT_t|P_t) \cdot P(PP_t|PT_{t-1}, PP_{t-1}, WT_{t-1}) \\ &\qquad \cdot P(WT_t|W_t, PP_t, PT_t) \cdot P(W_t|W_{t-1}, WT_{t-1}) \end{aligned} \tag{1}$$

The different conditional probability distributions (CPD) are defined as follows.

- Feature O. The observation feature O_t is a random function of the phone P_t in the CPD $P\left(O_t|P_t\right)$, which is denoted by a Gaussian Mixture Model as

  $$b_{P_t}(O_t) = P\left(O_t|P_t\right) = \sum_{k=1}^{M} \omega_{P_t k} N\left(O_t, \mu_{P_t k}, \sigma_{P_t k}\right) \tag{2}$$

  where $N\left(O_t, \mu_{P_t k}, \sigma_{P_t k}\right)$ is the normal distribution with mean $\mu_{P_t k}$ and covariance $\sigma_{P_t k}$, and $\omega_{P_t k}$ is the weight of the probability from the k^{th} mixture.
- Phone node P. The CPD $P(P_t|PP_t, W_t)$ is a deterministic function of its parents nodes phone position PP and word W :

  $$\begin{aligned} &P(P_t = j|W_t = i, PP_t = m) \\ &= \begin{cases} 1 & \textit{if} \;\; j \text{ is the } m-th \text{ phone of the word } i \\ 0 & \textit{otherwise} \end{cases} \end{aligned} \tag{5}$$

  This means that, given the current word W and the phone position PP, the phone P is known with certainty. For example, given the phone position PP as 2 in the word "five", we can know exactly the corresponding phone unit "ay".
- Phone transition probability PT which describes the probability of the transition from the current phone to the next phone. The CPD $P(PT_t|P_t)$ is a random distribution since each phone has a nonzero probability for staying at the current phone of a word or moving to the next phone.
- Phone position PP. It has three possible behaviors. (i) It might not change if the phone unit is not allowed to transit ($PT_{t-1} = 0$); (ii) It might increment by 1 if there is a phone transition ($PT_{t-1} = 1$) while the phone doesn't reach the last phone of the current word, i.e. the word unit doesn't transit ($WT_{t-1} = 0$); (iii) It might be reset to 0 if the word transition occurs ($WT_{t-1} = 1$).

$$P(PP_t = j \mid PP_{t-1} = i, WT_{t-1} = m, PT_{t-1} = n)$$
$$= \begin{cases} 1 & m = 1, j = 0 \\ 1 & m = 0, n = 1, j = i + 1 \\ 1 & \mathrm{m} = 0, n = 0, j = i \\ 0 & otherwise \end{cases} \tag{7}$$

- Word transition probability *WT*. In this model, each word is composed of its corresponding phones. The CPD $P(WT_t | W_t, PP_t, PT_t)$ is given by:

$$P(WT_t = j | W_t = a, PP_t = b, PT_t = m)$$
$$= \begin{cases} 1 & j = 1, m = 1, b = lastphone\,(a) \\ 1 & j = 0, m = 1, b =\sim lastphone\,(a) \\ 0 & otherwise \end{cases} \tag{8}$$

The condition $b = lastphone(a)$ means that *b* corresponds to the last phone of the word *a*, where *b* is the current position of the phone in the word. Equation (8) means that when the phone unit reaches the last phone of the current word, and phone transition is allowed, the word transition occurs with $WT_t = 1$.

- Word node *W*. In the training model, the word units are known from the transcriptions of the training sentences. In the decoding model, the word variable W_t uses the switching parent functionality, where the existence or implementation of an edge can depend on the value of some other variable(s) in the network, referred to as the switching parent(s). In this case, the switching parent is the word transition variable. When the word transition is zero ($WT_{t-1} = 0$), it causes the word variable to copy its previous value, i.e., $W_t = W_{t-1}$ with probability one. When a word transition occurs, $WT_{t-1} = 1$, however, it switches the implementation of the word-to-word edge to use bigram language model probability i.e. *bigram* which means the probability of one word transiting to another word whose value comes from the statistics of the training script sentences. The CPD $P(W_t = j \mid W_{t-1} = i, WT_t = m)$ is:

$$P(W_t = j \mid W_{t-1} = i, WT_t = m)$$
$$= \begin{cases} bigram(i, j) & if \quad m = 1 \\ 1 \quad if \quad m = 0, \quad i = j \\ 0 & otherwise \end{cases} \tag{9}$$

- In the training DBN model, the Word Counter (*WC*) node is incremented according to the following CPD:

$$p(WC_t = i \mid WC_{t-1} = j, WT_{t-1} = k, SS = l) =$$
$$\begin{cases} 1 & i = j \quad and \quad k = 0 \\ 1 & i = j \quad and \quad bound(w,j) \quad and \quad k = 1 \\ 1 & i = j+1 \quad and \quad \sim bound(w,j) \quad and \quad l = 0 \quad and \quad k = 1 \\ 1 & i = j+2 \quad and \quad \sim bound(w,j) \quad and \quad realword(w) \quad and \quad l = 1 \quad and \quad k = 1 \\ 1 & i = j+1 \quad and \quad \sim bound(w,j) \quad and \quad l = 1 \quad and \quad \sim realword(w) \quad and\ k = 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where $bound(w, j)$ is a binary indicator specifying if the position *j* of the current word *w* exceeds the boundary of the training sentence, if so, $bound(w, j) = 1$. $realword(w) = 1$ means that the coming word *w* after silence is a word with real meaning. If there is no word transition, $WC_t = WC_{t-1}$. On the contrary, if the word transits, the word number counts in different ways depending if the position of the word exceeds the boundary of the sentence: (i) if it does, word counter keeps the same as $WC_t = WC_{t-1}$; (ii) otherwise, it needs to check further the coming word, if there is no Skip Silence (SS) before the coming word, the word counter increments by one; If there is a SS, then check if the coming word has a real meaning, the word counter increments by 2 for the answer "yes" , and 1 for the answer "no".

3.2 Multi-Stream Asynchronous DBN Model (MSADBN)

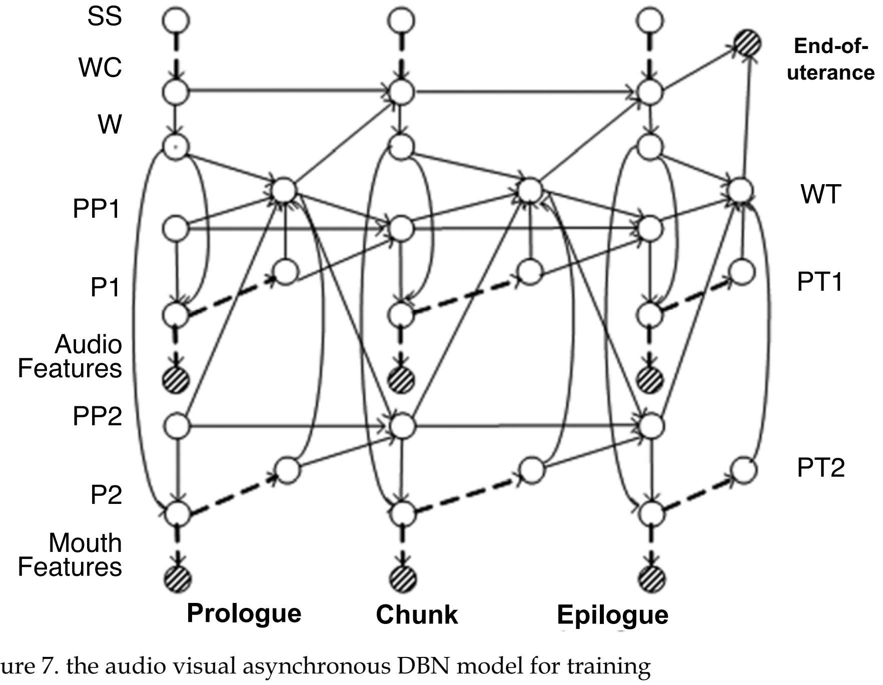


Figure 7. the audio visual asynchronous DBN model for training

For the audio visual speech unit segmentation, a multi-stream asynchronous DBN model is designed as shown in Fig.7. Audio features and visual features are imported into two independent graphical models in the same structure as the single stream DBN model of Fig.6, but are forced to be synchronized at the timing boundaries of words, by sharing one word variable *W* and word transition variable *WT*. In order to represent in the best way how the audio and visual speech are composed, between the succeeding word variables, the graphical topology of the audio stream with the phone variable *P1* is also allowed to be different with that of the visual stream with another phone variable *P2*. In other words, for each stream an independent phone variable is assigned, and its transition depends only on the phone instances inside the stream. This makes sure that the phone units in the audio stream and the visual stream may vary asynchronously. But at the same time, this asynchrony is limited inside the word by the word transition variable *WT*. The transition between words has to consider the cues from both the audio stream and the visual stream. Vice versa, it will affect the variations of both the phone positions *PP1* in the audio stream, and *PP2* in the visual stream: when and only when both the phone units in the two streams arrive the last phone position of a word, and both the phones are allowed to transit, the word transition occurs with the probability of 1. On the other hand, when a word transits to another word, it will reset the phone positions in the audio and in the visual stream to their initial values 0, in order to count the phone positions in the new word.

For the node variables that are not shared by the audio stream and visual stream as *W* and *WT*, the definitions of their conditional probability distributions are kept the same as that in the single stream DBN model. However, in this multi-stream DBN model, *WT* has five parent nodes: word *W*, phone position *PP1* and phone transition *PT1* in the audio stream, phone position *PP2* and phone transition *PT2* in the visual stream. As is explained above, the conditional probability distribution of *WT* can be defined as

$$\begin{aligned} & p(WT_t = j \mid W_t = a, PP1_t = b, PP2_t = c, PT1_t = m, PT2_t = n) \\ & = \begin{cases} 1 & j = 1, m = 1, n = 1, b = lastphone\,1(a), c = lastphone\,2(a) \\ 0 & otherwise \end{cases} \end{aligned} \tag{11}$$

where $b = lastphone1(a)$ and $c = lastphone2(a)$ mean that the phone position *b* in the audio stream, and the phone position *c* in the visual stream correspond to the last phone of the word *"a"*.

4. Experiments on Audio and Visual Speech

GMTK [Bilmes 2002] has been used for the inference and learning of the DBN models. In the experiments, we recorded our own audiovisual speech database with the scripts of the Aurora 3.0 audio database containing connected digits sequence from telephone dialing[Bilmes 2001]. 100 recorded sentences are selected as training data, and another 50 sentences as testing data. White noise with signal to noise ratio (SNR) ranging from 0dB to 30dB has been added to obtain noisy speech.

$$SNR = 10\log(E_s / E_N) \tag{12}$$

with E_s denoting the average energy of the speech sentence, and E_N the mean energy of white noise.

In the training process of the SDBN model, we first extract a word-to-phone dictionary from the standard TIMITDIC dictionary [Garofolo 1993] for the 11 digits. Actually, only 24 phone units--22 phones for the 11 digits together with the silence and short pause 'sp', are used due to the small size of the vocabulary. Associating each of the 22 phones with the observation features, the conditional probability distribution is modelled by 1 Gaussian mixture. Together with the silence model, and the short pause model which ties its state with the middle state of the silence model, a total parameter set of 25 Gaussian mixtures need to be trained.

For the MSADBN model, phone units are adopted in both graphical structures corresponding with audio and visual input. For each individual stream, the setup of the parameters is the same as that in the SDBN model. Therefore, for the 11 digits, 25 Gaussian mixtures are trained for the audio speech, and another 25 Gaussian mixtures for the visual speech.

To evaluate the speech recognition and phone segmentation performance of the SDBN and the MSADBN model, experiments are also done on the tied-state triphone HMMs with 8 Gaussian mixtures trained by the HMM toolkit HTK [Young 2006].

4.1 Speech Recognition Results

For the SDBN, Table 1 summarizes the word recognition rates (WRR) using acoustic features MFCC or PLP with white noise at different SNRs. Compared to the trained triphone HMMs, one can notice that with the SDBN model we obtain equivalent results in case of 'clean' signal and better results with strong noise. Over all (SNR=0dB to 30dB), SDBN with MFCC features shows an average improvement of 5.33%, and even 23.59% with PLP features in word accuracies over the triphone HMMs. Another interesting aspect of these results is that the improvement in word accuracies is more pronounced in cases of low SNRs.

Setup	0db	5db	10db	15db	20db	30db	Clean	0-30db
SDBN (MFCC_D_A)	42.94	66.10	71.75	77.97	81.36	96.61	97.74	72.79
SDBN (PLP_D_A)	76.27	88.70	92.09	93.79	97.18	98.31	98.87	91.05
Triphone HMM (MFCC_D_A)	27.55	42.76	58.67	83.85	93.82	98.10	98.34	67.46

Table 1. Speech recognition rate from audio stream (%)

For the speech recognition on the visual stream with SDBN, the word recognition rate is 67.26 percent for all SNR levels, which is also higher than 64.2% from the triphone HMMs.

Setup	0db	5db	10db	15db	20db	30db	Clean	0-30db
MSADBN (MFCC_D_A +GF)	53.94	70.61	86.06	89.39	93.03	95.76	97.27	81.46
MSADBN (PLP_D_A+GF)	90.96	94.92	96.05	96.61	97.19	98.87	98.31	95.76
MSHMM (MFCC_D_A +GF)	42.73	54.24	67.88	77.27	85.15	91.21	92.73	69.74

Table 2. Audio visual speech recognition rate (%)

Table 2 shows the word recognition rates from the audio visual multi-stream models. Comparing with the recognition results in Table 1 from only the audio stream, one can notice that in noisy environment, visual speech information, such as geographical features (GF) of lip, helps to improve the perception of speech. For the MSHMM, the MSADBN with MFCC features and MSADBN with PLP features, the average WRR improvements of 2.28%, 8.67% and 4.71% are obtained respectively for SNR=0dB to 30dB. Comparing the results from MSADBN and MSHMM with the same MFCC and geographical features, it can be seen that the designed MSADBN model outperforms MSHMM in modelling the dynamics of words in the two streams.

4.2 Phone Segmentation Results in the Audio Stream

Besides the word recognition results, the novelty of our work lies in the fact that we also obtain the phone segmentation sequence from the SDBN model, and further more, the asynchronous phone segmentation results in both audio and visual stream simultaneously from the MSADBN model.

Here we first evaluate the phone segmentation accuracies in the audio stream. An objective evaluation criterion, the phone segmentation accuracy (PSA) is proposed as follows: the phone segmentation results of the testing speech from the triphone HMMs are used as references. We convert the timing format of all the phone sequences obtained from the triphone HMM, the SDBN model and the MSADBN model with 10ms as frame rate, and then compare the phone units frame by frame. For each frame, if the segmented phone result from the SDBN (or MSADBN) model is the same as that in the reference, the score A is incremented. For the phone sequence of a speech sentence with C frames, the PSA is defined as

$$PSA = A / C \tag{11}$$

This evaluation criterion is very strict since it takes into account both phone recognition results together with their timing boundary information.

The average *PSA* values for the 50 testing sentences from the SDBN model and the MSADBN model with different features are shown in Table 3. One can notice that the SDBN model, either with MFCC features or with PLP features, gives phone segmentation results very close to those of the triphone HMMs, the standard continuous speech recognition models. While for the MSADBN model, the phone segmentations in the audio stream with timing boundaries differ even more to those from the triphone HMMs. This is reasonable because in the MSADBN model, it takes into account the phone units in the audio and in the visual streams simultaneously, and forces them to be synchronized on the timing boundaries of words.

As for the segmented timing boundaries of phones in the audio stream, whether the results from the single stream DBN model, or the results from the MSADBN model which also considers the visual features are more reasonable, will be discussed further in section 4.4.

Setup	0dB	5dB	10dB	15dB	20dB	30dB	Clean
SDBN (MFCC_D_A)	33.1	41.3	45.5	53.2	60.2	79.6	81.5
SDBN (PLP_D_A)	55.1	61.8	64.2	71.5	78.5	79.7	81.7
MSADBN (MFCC_D_A + GF)	25.4	26.9	28.2	31.4	33.3	38.4	40.4
MSADBN (PLP_D_A+GF)	35.2	36.0	39.2	40.8	42.1	45.6	46.4

Table 3. Phone segmentation accuracy in the audio stream (%)

4.3 Viseme Segmentation Results in the Visual Stream

To evaluate the phone segmentation results in the visual stream, the phone sequences are mapped to viseme sequences with a phone-to-viseme mapping table containing 16 viseme units that we previously proposed[Xie 2003]. The reference viseme sequences for the 50 testing sentences are obtained by manually labelling the image sequences, while also listening to the accompanying audio.

Viseme Segmentation Accuracy Firstly, the average viseme segmentation accuracies (VSA) for the 50 testing sentences from SDBN and MSADBN, calculated in the same way as PSA, are obtained as shown in Table 4. One can notice that MSADBN gets the improvement of 17.6% over SDBN for clean speech, with the consideration of audio cues. So by combining the audio and visual information together, we obtain more correct and accurate viseme segmentation in the visual speech.

Setup	0dB	5dB	10dB	15dB	20dB	30dB	Clean
SDBN (MFCC_D_A)	50.7	50.7	50.7	50.7	50.7	50.7	50.7
MSADBN (MFCC_D_A +GF)	51.5	52.9	56.2	58.5	59.8	62.5	68.3

Table 4. Viseme segmentation accuracy (%)

Relative Viseme Segmentation Accuracy (RVSA) VSA is a very strict criterion for evaluating the viseme segmentation results in the visual stream: only when the corresponding viseme units are correctly recognized, and the segmented timing boundaries are the same with the references, the score can be incremented. On the other hand, in our task, the purpose of segmenting the viseme sequence is for constructing a mouth animation in the future: suppose for each viseme, one representative mouth image is built, a mouth animation can be constructed by concatenating the mouth images corresponding to the viseme sequence. So from the speech intelligibility point of view, it is also acceptable if a viseme is recognized as another viseme unit with very similar mouth shape. RVSA is thus a measurement to evaluate the global similarity between the mouth animation constructed from the reference viseme sequence, and the one constructed from the segmented viseme sequence.

The RVSA is calculated as follows:

For two gray level images I_{ref} and I with the same size, their difference can be measured by Mean Square Error

$$MSE(I_{ref},I)=\frac{1}{M}\sum_u\sum_v(I_{ref}(u,v)-I(u,v))^2 \tag{12}$$

with M the number of pixels, $I(u,v)$ the gray level of the pixel on the coordinate (u,v). The difference of an arbitrary viseme representative image pair (vis_m,vis_n) can be measured by eq. (12) and then be normalized to a viseme similarity weight (VSW) which shows their similarity contrariwise.

$$VSW(vis_m,vis_n)=1-\frac{MSE(vis_m,vis_n)}{\underset{i,j=1,2,...N}{MAX}\left(MSE(vis_i,vis_j)\right)} \tag{13}$$

The lower the difference is, the higher the VSW is. If two visemes are totally the same, the VSW is 1.

Therefore, for the whole viseme image sequence of *N* frames, the RVSA can be calculated by the VSW between the resulting and reference visemes on each image frame t.

$$RVSA=\frac{1}{N}\sum_t VSW_t \tag{14}$$

Table 5 shows the average RVSA values of the viseme sequences for the 50 testing sentences, obtained from the SDBN model and from the MSADBN respectively. One can notice that in relative clean environment with SNR higher than 20dB, the MSADBN model creates more close mouth shape animations to the reference animations. While with the increasing of noise, the viseme segmentations from MSADBN will get worse.

Setup	0dB	5dB	10dB	15dB	20dB	30dB	Clean
SDBN	75.1	75.1	75.1	75.1	75.1	75.1	75.1
MSADBN (MFCC_D_A+GF)	71.8	73.1	73.8	74.3	75.6	77.6	82.6

Table 5. Relative viseme segmentation accuracy based on image (%)

4.3 Asynchronous Segmentation Timing Boundaries

The asynchronous timing boundaries of phone segmentation in the audio stream, and the viseme segmentation in the visual stream can be obtained, either by performing the SDBN for two times (with audio features and with visual features respectively), or by inputting the audio and visual features into the MSADBN simultaneously. As an example to illustrate the segmentation results, we adopt a simple audio-video speech corresponding to the sentence "two nine". Table 6 shows the segmentation results in both audio and visual streams, together with the manually labeled phone and viseme timing boundaries as references. To see clearly the asynchrony between the phones and visemes obtained from the SDBN model and the MSADBN model, the results are also expressed in Fig.8, together with the temporal changes of audio and visual features. The mapping relationships between visemes and phones are: viseme 'vm' corresponds to the mouth shape of the phone 't', 'vg' to phone 'uw', 'vp' to phone 'n', and 'vc' to phone 'ay'.

setup	t (vm)	uw (vg)	n (vp)	ay (vc)	n (vp)
Reference (audio)	0.43-0.56	0.56-0.83	0.83-0.98	0.98-1.23	1.23-1.39
Reference(visual)	0.36-0.44	0.45-0.76	0.77-0.88	0.89-1.20	1.21-1.36
SDBN(audio)	0.43-0.54	0.55-0.84	0.85-0.97	0.98-1.20	1.21-1.34
SDBN(visual)	0.36-0.40	0.41-0.67	0.68-0.71	0.72-1.18	1.19-1.24
MSADBN(audio)	0.42-0.50	0.51-0.76	0.77-0.99	1.00-1.29	1.30-1.36
MSADBN(visual)	0.42-0.45	0.46-0.76	0.77-0.87	0.88-0.97	0.98-1.36

Table 6. The phone and viseme segmentation timing boundaries (s)

From Fig.8, one can notice that for the single stream segmentation from SDBN, without considering the mutual effect of the other accompanying cue, the asynchronies between the phones and the corresponding visemes are quite large. For example, the first viseme "vm" in the visual stream ends at 0.54s, preceding about 140ms than the corresponding phone unit "t" in the audio stream. And for the viseme "vp" (phone "n") at the beginning of the word "nine", the ending time in the visual stream is even 260ms earlier than that in the audio stream.

Considering the asynchronous segmentation from the MSADBN model, one can notice that by integrating the audio features and visual features in one model, and forcing them to be aligned at the timing boundaries of words, that for most of the segmented timing boundaries in the two streams, they seem to attract each other to be closer. Now the distance of the ending times in the visual stream and in the audio stream, becomes 50ms for the first viseme "vm" (phone "t"), and 120ms for the first viseme "vp" (phone "n") of "nine". But this behavior is not always this case, for example, an exception occurs with the ending times of the viseme "vc" (phone "ay"). From the MSADBN model, the distance between the timing boundaries in visual and in audio stream is 220ms, which is longer than 20ms obtained from the SDBN model. Comparing the asynchronous audio visual phone (viseme) segmentation results from the SDBN model and the MDBN model with the reference timing boundaries, we can see that in most cases, the asynchrony obtained from the MSADBN model is more reasonable.

5. Discussion

In this chapter, we first implement an audio or visual single stream DBN model proposed in [Bilmes 2005], which we demonstrate that it can break through the limitation of the state-of-the-art 'whole-word-state DBN' models and output phone (viseme) segmentation results. Then we expand this model to an audio-visual multi-stream asynchronous DBN (MSADBN) model. In this MSADBN model, the asynchrony between audio and visual speech is allowed to exceed the timing boundaries of phones/visemes, in opposite to the multi-stream hidden markov models (MSHMM) or product HMM (PHMM) which constrain the audio stream and visual stream to be synchronized at the phone/viseme boundaries.

In order to evaluate the performances of the proposed DBN models on word recognition and subunit segmentation, besides the word recognition rate (WRR) criterion, the timing boundaries of the segmented phones in the audio stream are compared to those obtained from the well trained triphone HMMs using HTK. The viseme timing boundaries are compared to manually labeled timing boundaries in the visual stream. Furthermore, suppose for each viseme, one representative image is built and hence a mouth animation is constructed using the segmented viseme sequence, the relative viseme segmentation accuracy (RVSA) is evaluated from the speech intelligibility aspect, by the global image sequence similarity between the mouth animations obtained from the segmented and the reference viseme sequences. Finally, the asynchrony between the segmented audio and visual subunits is also analyzed. Experiment results show: 1) the SDBN model for audio or visual speech recognition has higher word recognition performance than the triphone HMM, and with the increasing noise in the audio stream, the SDBN model shows more robust tendency; 2) in a noisy environment, the MSADBN model has higher WRR than the SDBN model, showing that the visual information increases the intelligibility of speech. 3) compared with the segmentation results by running the SDBN model on audio features and on visual features respectively, the MSADBN model, by integrating the audio features and visual features in one scheme and forcing them to be synchronized on the timing boundaries of words, in most cases, gets more reasonable asynchronous relationship between the speech units in the audio and visual streams.

In our future work, we will expand the MSADBN model to the subunits segmentation task of a large vocabulary audio visual continuous speech database, and test its performance in speech recognition, as well as analyze its ability of finding the inherent asynchrony between audio and visual speech.

6. Acknowledgement

This research has been conducted within the "Audio Visual Speech Recognition and Synthesis: Bimodal Approach" project funded in the framework of the Bilateral Scientific and Technological Collaboration between Flanders, Belgium(BILO4/CN/02A) and the Ministry of Science and Technology (MOST), China([2004]487), and the fund of 'The Developing Program for Outstanding Persons' in NPU—DPOP: NO. 04XD0102.

7. References

Bilmes, J. & Zweig, G. (2001). *Discriminatively structured dynamic graphical models for speech recognition*. Technical report, JHU 2001 Summer Workshop, 2001.

Bilmes, J., Zweig, G.: The graphical models toolkit: an open source software system for speech and time-series processing. *Proceedings of the IEEE Int. Conf. on Acoustic Speech and Signal Processing(ICASSP)*. Vol. 4(2002), pp. 3916-3919.

Bilmes,J. & Bartels, C. (2005). Graphical Model Architectures for Speech Recognition. *IEEE Signal Processing Magazine*, vol.22, pp. 89-100, 2005.

Eisert, P. (2000). *Very low bit-rate video coding using 3-d models*. PhD thesis, Universitat Erlangen, Shaker Verlag, Aachen, Germany (2000) , ISBN 3-8265-8308-6.

Garofolo, J.S.; Lamel, L.F & Fisher, W.M. et al (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus, *Linguistic Data Consortium*, Philadelphia. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1.

Gowdy, J.; Subramanya, A.; Bartels, C. & Bilmes, J. (2004). DBN-based multistream models for audio-visual speech recognition. *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 993–996, May 2004.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*. Vol. 87, Issue 4 (1990), pp. 1738-1752.

Lee, Y.; Terzopoulos, D. & Waters, K. (1993). Constructing physics based facial models of individuals. *Proceedings of the Graphics Interface '93 Conference*, Toronto, Canada. (1993), pp. 1-8.

Ravyse, I. (2006). *Facial Analysis and Synthesis*. PhD thesis, Vrije Universiteit Brussel, Dept. Electronics and Informatics, Belgium. online: www.etro.vub.ac.be/Personal/icravyse/RavysePhDThesis.pdf.

Ravyse, I.; Enescu, V. & Sahli, H. (2005). Kernel-based head tracker for videophony. *The IEEE international Conference on Image Processing 2005 (ICIP2005)*, Vol. 3, pp.1068-1071, Genoa, Italy, 11-14/09/2005.

Steven, B.D.& P.M. (1980). Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing. Vol. 28, (1980) pp. 357-366.

Vedula, S. & Baker, S. et al (2005). Three-dimensional scene flow. *IEEE transactions on Pattern Analysis and Machine Intelligence*. Vol. 27, (2005), pp. 137-154.

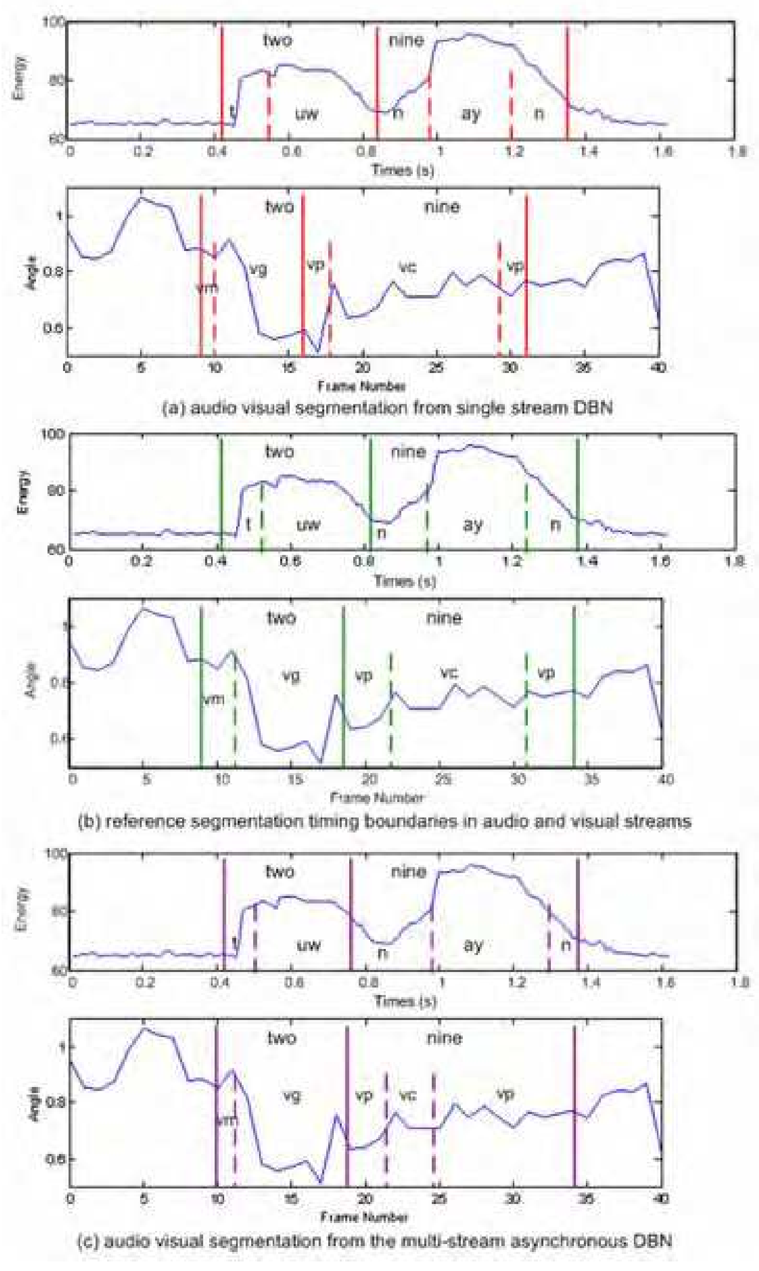
Xie, L. & Jiang, D.M.et al (2003). Context dependent viseme models for voice driven animation. *4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications, EC-VIP-MC 2003*, Vol.2 (2003), pp. 649-654, Zagreb, Croatia, July, 2003.

Young, S.J.; Kershaw, D.; Odell, J. & Woodland, P. (2006). The HTK Book (for HTK Version 3.4). http://htk.eng.cam.ac.uk/docs/docs.shtml.

Zhang, Y.; Diao, Q., & Huang, S. et al (2003). DBN based multi-stream models for speech. *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 836–839, Hong Kong, China, Apr. 2003.

Zhou, Y.; Gu, L. & Zhang, H.J. (2003). Bayesian tangent shape model: Estimating shape and pose parameters via bayesian inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR2003)*. Vol. 1. pp. 109-118, 2003.

Zweig, G. (1998). *Speech recognition with dynamic Bayesian networks*, Ph.D. dissertation, Univ. California, Berkeley,1998.

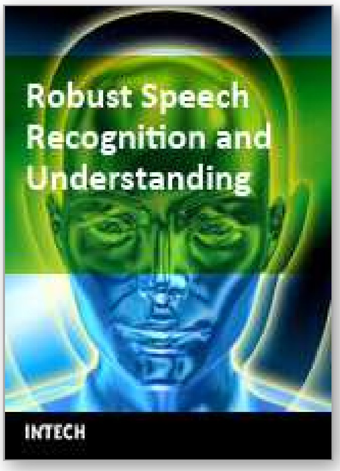


(a) audio visual segmentation from single stream DBN

(b) reference segmentation timing boundaries in audio and visual streams

(c) audio visual segmentation from the multi-stream asynchronous DBN

Figure 8. Audio visual asynchronous phone/viseme segmentation results



Robust Speech Recognition and Understanding
Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821