16

Multi-Stream Asynchrony Modeling for Audio Visual Speech Recognition

Guoyun Lv, Yangyu Fan, Dongmei Jiang and Rongchun Zhao
NorthWestern Polytechnical University (NWPU)
127 Youyi Xilu, Xi'an 710072,
P.R.China

1. Introduction

The success of currently available speech recognition systems was restricted to relative controlled laboratory conditions or application fields. The performance of these systems rapidly degraded in more realistic application environments (Lippmann, 1997). Since the vast majority of background noise were introduced by transmission channel, microphone distance or environment noise, some new audio feature extraction methods (Perceptual Linear Predictive (PLP), RelAtive SpecTrAl (RASTA) (Hermansky, 1990; Hermansky,1994), and other ways such as vocal tract length normalization and parallel model combination for speech and noise, were used to describe the complex speech variations. Though these methods improved the system robustness to noisy environment to some extent, but their advantage was limited.

Since both human speech production and perception are bimodal in nature (Potamianos et al, 2003), visual speech information from the speaker's mouth has been successfully shown to improve noisy robustness of automatic speech recognizers (Dupont & Luettin 2000; Gravier et al, 2002). There are two main challenging problems in the reported Audio-Visual Speech Recognition (AVSR) systems (Nefian et al, 2002; Gravier et al, 2002): First, the design of the visual front end, i.e. how to obtain the more static visual speech feature; second, how to build a audio-visual fusion model that describes the inherent correlation and asynchrony of audio and visual speech. In this paper, we concentrate on the latter issue.

Previous works on combining multiple features can be divided into three categories: feature fusion, decision fusion and model fusion. Model fusion seems to be the best technique to integrate information from two or more streams. However, the experiments results of many AVSR systems show that although the visual activity and audio signal are correlative, but they are not synchronous, the visual activity often precedes the audio signal about 120ms (Gravier et al, 2002; Potamianos et al, 2003) . Each AVSR system should take the asynchrony into account.

Since hidden Markov model (HMM) based models achieve promising performance in speech recognition, many literatures have adopted Multi-Stream HMM (MSHMM) to integrate audio and visual speech feature (Gravier et al, 2002; Potamianos et al, 2003; Nefian et al, 2002), such as State Synchrony Multi-Stream HMM (SS-MSHMM), State Asynchrony Multi-Stream HMM (SA-MSHMM) (Potamianos et al, 2003), Product HMM (PHMM) (Dupont, 2000), Couple HMM (CHMM) and Factorial HMM (FHMM) (Nefian et al, 2002)

Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert,
ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria

and so on. In these models, audio and visual features are imported to two or more parallel HMMs with different topology structures respectively, but on some nodes, such as phone, syllable et al; some constraints are imposed to limit the asynchrony of audio and visual streams to state (phone or syllable) level. These MSHMMs describe the correlation and asynchrony of audio and visual speech to some extent. Compared with the single stream HMM, system performance is improved especially in noisy speech environment, but these MSHMMs can only use phone as recognition unit for speech recognition task on a middle or large vocabulary audio-visual database. It constrains the audio and visual stream to be synchronous in the phonemic boundary. However, the asynchrony of audio and visual stream exceeds phonemic boundary in many conditions. The better recognition rate should be obtained if loosing the asynchrony limitation of audio and visual stream.

In recent years, it was an active research topic to adopt Dynamic Bayesian Network (DBN) for speech recognition (Bilmes, 2002; Murphy, 2002; Zweig, 1998). DBN model is a statistic model that can represent multiple collections of random variables as they evolve over time. It is appropriate to describe complex variables and conditional relationship among the variables, since it can automatically learn the conditional probability distribution among the variables, with better extensible performance. Bilmes, Zweig et al, used single stream DBN model for isolated words and small vocabulary speech recognition (Bilmes et al, 2001; Lv et al, 2007). Zhang YM proposed a multi-stream DBN model for speech recognition by combining different audio features (MFCC, PLP, RASTA) (Zhang et al, 2003), although the model described the asynchrony of audio and visual streams by sharing the same word node, while in fact, there are not asynchrony for different audio features from the same voice. N. Gowdy expanded this model for audio-visual speech recognition (Gowdy et al, 2003), an improvement was obtained in word accuracy, while between the word nodes, and each stream is not complete independence, which affected the asynchrony of both streams to some extent. Bimes proposed a general multi-stream asynchrony DBN model structure (Bilmes & Bartels, 2005), in this model, the word transition probability is determined by the state transitions and the state positions both in the audio stream and in the visual stream. Between the word nodes, two streams have their own nodes and the dependent relationship between the nodes. But no more experimental results were given.

In this work, we use the general multi-stream DBN model structure given in (Bilmes & Bartels, 2005) as our baseline model. In (Bilmes & Bartels, 2005), both in audio stream and in visual stream, each word is composed of the fixed number of states, and each state is associated with observation vectors. The training parameters are very tremendous, especially for the task of large vocabulary speech recognition. In order to reduce the training parameters, in our model, both in audio stream and in visual stream, each word is composed of its corresponding phones sequence, and each phone is associated with observation vector. Since phones are shared by all the words, the training parameter will be enormously reduced, and we name it Multi-Stream Asynchrony DBN (MS-ADBN) model. But MS-ADBN model is word model whose recognition basic units are words. It is not appropriate for the task of large vocabulary AVSR. Base on MS-ADBN model, an extra hidden node level—state is added between phone node level and observation variable level in both stream, resulting in a novel Multi-stream Multi-states Asynchrony DBN (MM-ADBN) model. In MM-ADBN model, each phone is composed of fixed number of states, and each state is associated with observation vector, besides word, dynamic pronunciation process of phone is also described. Its recognition basic units are phones, and can be used for large vocabulary audio-visual speech recognition.

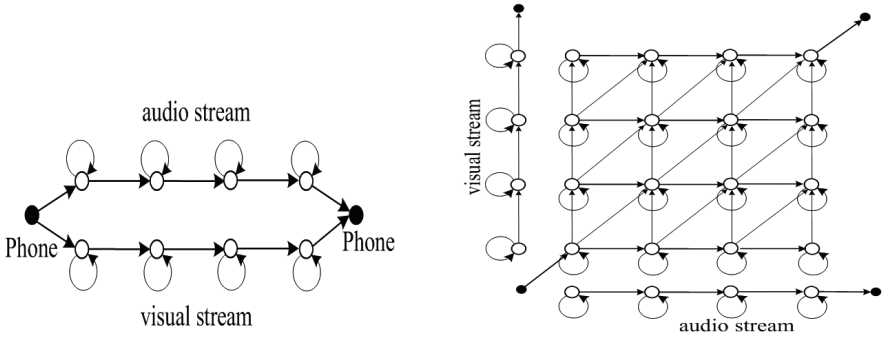
The paper is organized as follows. Section 2 describes the structures and conditional probability distributions of the proposed MS-ADBN model and MM-ADBN model. In section 3, experiments and evaluations are given, followed by our conclusions in section 4.

2. Multi-stream asynchrony model

In this section, at first, we briefly review the previous MSHMM, and then we describe the multi-stream asynchrony DBN model proposed in our work. Finally, we make simple comparisons for these audio-visual speech recognition models.

2.1 State asynchrony multi-stream HMM

Multi-Stream hidden Markov model (MSHMM) was a popular method within audio-visual model fusion framework. The MSHMM linearly combines the class log-likelihoods based on the audio-only and video-only observations at a number of possible stages (such as state, phone et al). In early most cases, the synchronous point of audio stream and visual stream is at the HMM state level, and we name it State Synchronous MSHMM (SS-MSHMM). To take asynchrony of audio and visual stream into account, the synchronous point should be taken to a coarser level, such as phone, syllable, or word level. However, on one hand, for middle and large vocabulary speech recognition, the phone recognition unit must be used; on the other hand, to implement easily, previous popular works often use the state asynchrony multi-stream HMM (SA-MSHMM) (Gravier, 2002; Nefian et al, 2002), and the synchronous points are taken to the phonemic boundaries. Because of the limitation of HMM expression ability, such a model can be implemented as a product HMM (PHMM), as illustrated in Fig. 1. Typically, SA-MSHMM with four audio and four video HMM states is given in Fig. 1 (a), and its corresponding product HMM is given in Fig. 1(b).



(a) SA-MSHMM with four audio and video states (b) Corresponding product HMM

Fig. 1. Illustration of SA-MSHMM and its corresponding product HMM

The observation probability can be described as:

$$b_j(o_t) = \prod_{s\in\{a,v\}} [\sum_{m=1}^{M_s} \omega_{jsm} N(o_t^s; \mu_{jsm}; \sigma_{jsm})]^{\lambda_s} \quad (1)$$

Where, $s = a, v$, respectively for audio and visual stream. λ_s is stream exponent, $\lambda_a + \lambda_v = 1$; j is state node, m means mth Gaussian mixture unit, M_s is number of Gaussian mixture unit

in s^{th} stream, ω_{jsm} denotes weight value, μ and σ is the mean and covariance of Gaussian distribution $N(.)$.

Although SA-MSHMM can describe the asynchrony of audio and visual stream to some extent, But problems remain due to the inherent limitation of the HMM structure. On one hand, for large vocabulary speech recognition tasks, phones are the basic modeling units, the model will force the audio stream and the visual stream 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. On the other hand, Once a little slight varieties are done on MSHMM, a large amount of human effort must be placed into making significant modifications on top of already complex software without having any guarantees about their performance. So a new and unified multi-stream model framework is expected to loose the limitation of asynchrony of audio stream and visual stream to the coarser level.

2.2 MS-ADBN model

A Dynamic Bayesian Network (DBN) is a statistical model that can represent collections of random variables and their dependency relationships as they evolve over time. HMM is just special case of much more general DBN model. Comparing with HMM, DBN model has a more flexible and extensible structure, and explicitly describes the hierarchical relationship of main components (e. g word, phone, state and observation) of speech recognition. In general, the DBN model meets two conditions: 1) except the initial frame, the topology structure is same in each frame; 2) the condition probability relationship between the frames follows the one-order Markov model. Additionally, Uniform training and decoding algorithm make the implement of DBN model become easier.

Since DBN model has some preponderant on describing the complex model structure, multi-stream DBN model is expected to model the audio-visual speech recognition structure by loosing the asynchrony of the audio stream and visual stream.

Fig. 2 illustrates the recognition structure of a multi-stream asynchrony DBN (MS-ADBN) model. It is composed of a Prologue part (initialization), a Chunk part that is repeated every time frame (t), and a closure of sentence with an Epilogue part. Abbreviation of every node is denoted in the parentheses: (W) is the word unit in a sentence; (WT) is the occurrence of a transition from one word to another word; (PP1) and (PP2) are the position of the current phone in the current word; (PT1) and (PT2) are the occurrence of a transition from a phone to another phone; (P1) and (P2) is the phone node; O1 is acoustic observation; O2 is visual observation vector. The nodes with shade are the observation variables, and the nodes without shades are the hidden state variables.

In MS-ADBN model, the word variable and word transition variable are at the top of the structure, when a word transition occurs, it will reset (PP1) and (PP2) to their initial value, hence audio stream and visual stream are forced to be synchronous in the same word node.

While between the word nodes, each stream has its own independence nodes and conditional probability distributions between the nodes, each word is composed of its corresponding composed phones, and each phone is associated with observation features. Namely, it allows two independence representations for dynamic pronunciation process of a word in this model. Additionally, word transition is determined by audio and visual steam together, to make word transition occur, we must have that both PP1 and PP2 are the last phone of current word, as well as both PT1 and PT2 occurs. Comparing with MSHMM, the asynchrony of audio and visual stream is really loosed to word level.

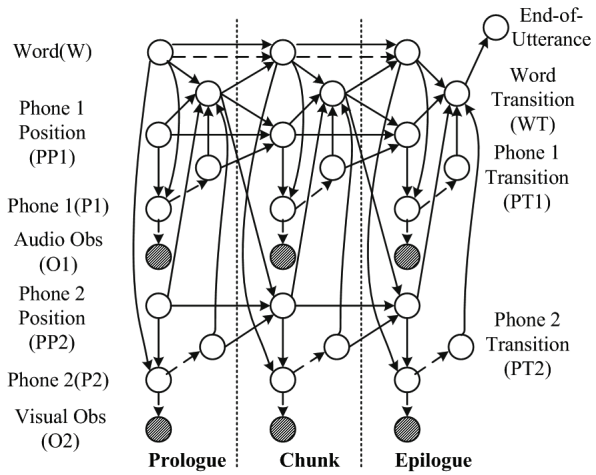


Fig. 2. MS-ADBN speech recognition model

While MS-ADBN model only describes the dynamic pronunciation process of word, it is a word model whose basic recognition units are words. It is only appropriate for small vocabulary speech recognition. For large vocabulary audio visual speech recognition, the less recognition sub-word units—phone, should be modeled.

2.3 MM-ADBN model

To describe the dynamic process of phone, we proposed a novel multi-stream multi-states asynchrony DBN (MM-ADBN) model given in Fig. 3.

It can be seen that MM-ADBN model is an augmentation of the MS-ADBN model, to which we add some extra hidden nodes (state, state position and state transition) and corresponding conditional probability relationships between phone variable level and observation variable level in both audio stream and visual stream, and new additive parts are labelled with red color in Fig.3. In MM-ADBN model, both in audio stream and in visual stream, each phone is composed of fixed number of states, and each state is associated with observation vector. Hence, it is a phone model whose basic recognition units are phones, and can be used for large vocabulary audio-visual speech recognition. In Fig. 3, the definitions of the word and phone related nodes are the same as those in the MS-ADBN model. The new notations: (SP1) and (SP2) are the position of the current state in the current phone; (ST1) and (ST2) are the occurrence of a transition from a state to another state, defined similarly as (PT1) or (PT2); (S1) and (S2) are the state node. Suppose the input speech contains T frames of features, and the set of all the hidden nodes is denoted as $H_{1:T}$.

$$\begin{aligned} H_{1:T} = (&W_{1:T}, WT_{1:T}, PP1_{1:T}, PP2_{1:T}, PT1_{1:T}, PT2_{1:T}, SP1_{1:T}, \\ &SP2_{1:T}, ST1_{1:T}, ST2_{1:T}, P1_{1:T}, P2_{1:T}, S1_{1:T}, S2_{1:T}) \end{aligned} \tag{2}$$

For the model given in Fig. 3, the probability of observation can be computed as

$$p(O1_{1:T}, O2_{1:T}) = \sum_{H_{1:T}} p(H_{1:T}, O1_{1:T}, O2_{1:T}) \tag{3}$$

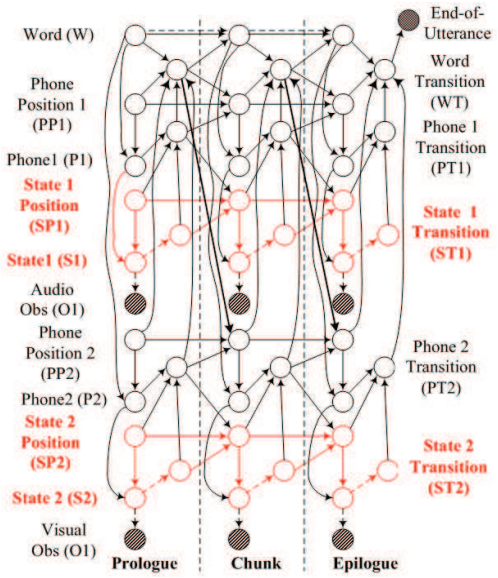


Fig. 3. MM-ADBN speech recognition model

For a better understanding of MM-ADBN model, we describe each node's Conditional Probability Distributions (CPD) as follows.

a. Observation Feature (O1 and O2). The observation feature Ox_t is a random function of the state Sx_t with the CPD $P(Ox_t|Sx_t)$, which is described by a normal Gaussian Model $N(O_t, \mu_{Sx_t}, \sigma_{Sx_t})$ with mean μ_{Sx_t} and covariance σ_{Sx_t}, where symbol x denotes audio stream or visual stream, x=1 means audio stream and x=2 means visual stream, which also be used in the following expression.

b. State transition probability (ST1 and ST2), which describe the probability of the transition from the current state to the next state in the audio stream and visual stream respectively. The CPD $P(STx_t|Sx_t)$ is random since each state has a nonzero probability for staying at the current state or moving to the next state, in the initial frame, the CPD is assumed as 0.5.

c. State node (S1 and S2), since each phone is composed of fixed number of states, giving the current phone and the position of the current state in the phone, the state Sx_t is known with certainty.

$$p(Sx_t = j | Px_t = i, SPx_t = m) = \begin{cases} 1 & if\ j\ \text{is the}\ m-th\ \text{state of the phone}\ i \\ 0 & otherwise \end{cases} \quad (4)$$

d. State position (SP1 and SP2), in the initial frame, the initial value is zero. In the other time slices, its CPD has three behaviors, (i) It might not change from one frame to the next frame when there is no state transition and phone transition; (ii) It might increment

by 1 when there is a state transition and the model is not in the last state of the phone; (iii) It might be reset to 0 when a phone transition occurs.

$$p(SPx_t = j \mid SPx_{t-1} = i, PTx_{t-1} = m, STx_{t-1} = n) = \begin{cases} 1 & m = 1, j = 0 \\ 1 & \text{m} = 0,\text{n} = 0,\text{j} = \text{i} \\ 1 & m = 0, n = 1, j = i + 1 \\ 0 & otherwise \end{cases} \tag{5}$$

e. Phone node (P1 and P2). Each word is composed of its corresponding phones, giving the current word and the position of the current phone in the word, the phone Px_t is known with certainty.

$$p(Px_t = j \mid W_t = i, PPx_t = m) = \begin{cases} 1 & j \text{ is the } m-th \text{ phone of the word } i \\ 0 & otherwise \end{cases} \tag{6}$$

f. Phone position (PP1 and PP2). In the initial frames, the initial value is zero. In the other time frame, the CPD is as follows.

$$p(PPx_t = j \mid PPx_{t-1} = i, WT_{t-1} = m, PTx_{t-1} = n) = \begin{cases} 1 & m = 1, j = 0 \text{ or } \text{m} = 0,\text{n} = 0,\text{j} = \text{i} \\ 1 & m = 0, n = 1, j = i + 1 \\ 0 & otherwise \end{cases} \tag{7}$$

g. Phone transition (PT1 and PT2). In this model, each phone is composed of fixed number of states. The CPD $P(PTx_t \mid Px_t, SPx_t, STx_t)$ is given by:

$$P(PTx_t = j \mid Px_t = a, SPx_t = b, STx_t = m) = \begin{cases} 1 & j = 1, m = 1, b = laststate(a) \\ 1 & j = 0, m = 1, b = \sim laststate(a) \\ 0 & otherwise \end{cases} \tag{8}$$

Where $laststate(a)$ denotes the last state of phone 'a'. Only when the last state of a phone is reached, and a state transition is allowed, the current phone can transit to a new phone unit.

h. Word transition (WT), which is determined by audio stream and visual stream together.

$$p(WT_t = j \mid W_t = a, PP1_t = b, PP2_t = c, PT1_t = m, PT2_t = n) = \begin{cases} 1 & j = 1, m = 1, n = 1, b = lastphone1(a), c = lastphone2(a) \\ 1 & j = 0 \ (m \neq 1 \text{ or } n \neq 1 \text{ or } b \neq lastphone1(a) \text{ or } c \neq lastphone2(a) \\ 0 & otherwise \end{cases} \tag{9}$$

The condition $b = lastphone1(a)$ means that b corresponds to the last phone of the word 'a' in the audio stream. Similarly, the condition $c = lastphone2(a)$ means that c corresponds to

the last phone of the word 'a' in the visual stream. Equation (9) means that when the phone units reach the last phone of the current word for both in audio stream and in visual stream respectively, and phone transitions for both two streams are allowed, the word transition occurs. Otherwise, word transition is not changed.

i. Word node (W), in initial frame, the word variable W starts out using a unigram distribution over words in the vocabulary. In the other frames, the word variable W_t depends on W_{t-1} and WT_t with CPD $P(W_t = j \mid W_{t-1} = i, WT_t = m)s$.

$$P(W_t = j \mid W_{t-1} = i, WT_t = m) = \begin{cases} bigram(i, j) & if \quad m = 1 \\ 1 & if \quad m = 0, \quad i = j \\ 0 & otherwise \end{cases} \tag{10}$$

where $bigram(i, j)$ means the transition probability from word i to word j.

2.4 Comparison of model

It can be known from the above models, main differences of several models are as follows.

1. The asynchrony of audio stream and visual stream: SS-MSHMM describes the asynchrony of the two streams at the HMM state level, and SA-MSHMM (be implemented as PHMM) looses the asynchrony to the phone boundary level, while MS-ADBN model and MM-ADBN model describe the asynchrony of both streams at word level.
2. The different recognition unit: MS-ADBN model is a word model whose recognition basic units are words. While MM-ADBN model is a phone model whose recognition basic units are phones, and can be used for large vocabulary audio visual speech recognition.

In the following experiments section, for the models proposed in this section, we will make some comparisons on the sake of different asynchrony description and different recognition basic unit.

3. Experiments and evaluation

In our work, The Graphical Models Toolkit (GMTK) (Bilmes & Zweig, 2002) has been used for the inference, learning and recognition of all the DBN models and Hidden Markov Model Toolkit (HTK) (Young, 1994) has been used for all HMMs. Speech recognition experiments have been done on a continuously spoken audio-visual digit database and an audio-visual large vocabulary continuous speech database respectively.

3.1 Audio visual database description

The digit continuous audio-visual database has been recorded with the scripts from the AURORA database 2.0 (Hirsch & Pearce 2000) which contains digit sequence from telephone dialing. Each sequence contains several digits from the digit set {zero, one, …, nine, oh}. 22 phone units are obtained by transcribing the digit set with the TIMIT dictionary. 100 clean audio-visual sentences have been selected as training set, and another 50 audio-visual sentences as testing set. White noise with signal to noise ratio (SNR) ranging from 0dB to 30dB has been added to obtain noisy speech.

The continuous audio-visual experiments database has been recorded with the scripts from the TIMIT database. 6 people's 600 sentences containing 1693 word units have been used in our experiments. Totally 76 phone units (including "silence" and short pause "sp") are obtained by transcribing the sentence scripts into phone sequences using the TIMIT dictionary. Since the database is relatively small for large vocabulary audio-visual speech recognition. To test performance of MM-ADBN model, we use the jackknife procedure, 600 sentences were split up in six equal parts, and six recognition experiments were carried out. In each recognition experiment, 500 sentences are used as training set, the remaining 100 sentences as testing set. Report test results are the average of the results of six experiments. While for MS-ADBN model, since it is word model, to avoid the case that some words in the testing sentence may not appear in the training set, all 600 sentences are used as training set and testing set. Noisy environments are also considered by adding white noise with SNRs ranging from 0dB to 30dB as testing set.

The above two databases are recorded with the same condition: with high-quality camera, clean speech environment, uniform background and lighting. The face of the speaker in the video sequence is high-quality frontal upright, and video is MPEG2-encoded at a resolution of 704×480, and at 25Hz.

3.2 Audio and visual feature extraction

Mel Filterbank Cepstrum Coefficients (MFCC) features are extracted by HTK with the frame rate of 100 frames/s. 13 MFCC features, energy, together with their delta and acceleration coefficient, resulting in a feature vector of 42 acoustic features (MFCC_E_D_A) has been extracted.

Visual feature extraction is given in Fig. 4, which starts with the detection and tracking of the speaker's face (Ravyse et al, 2006), Since the mouth is the most important speech organs, the contour of the lips is obtained through the Bayesian Tangent Shape Model (BTSM) (Zhou et al, 2003), for each image, 20 profile points include outer contour and inner contour of the mouth are obtained, which is given in Fig. 5. Based on these profile feature points, we extract a 20 dimensional geometrical feature vector: 5 vertical distance features and 5 horizontal distance features between outer contour feature points, 3 vertical distance features and 3 horizontal distance features between inner contour feature points, 4 angle features (α, β, θ and φ). Sketch map of the features are given in Fig. 5.

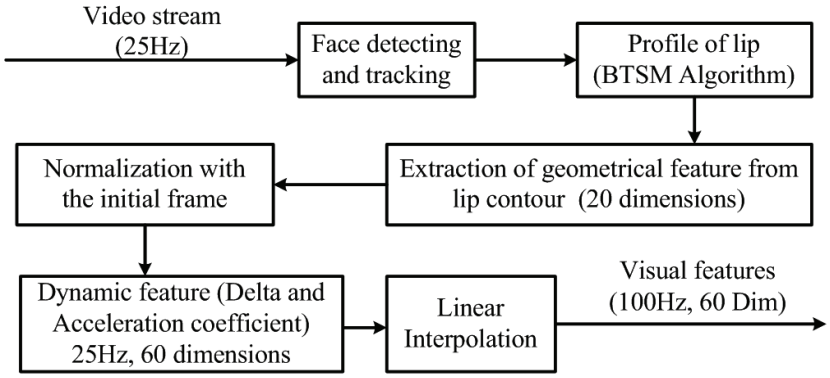
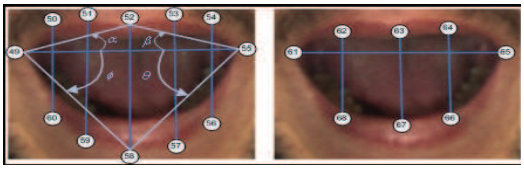


Fig. 4. Visual feature extraction



(a) Outer contour (b) Inner contour

Fig. 5. Profiles and geometrical features of lip

To eliminate the different speaker's affection, all visual features are normalized by subtracting the corresponding initial frame geometrical feature. In order to describe the dynamic visual feature, we extract the delta and acceleration coefficient of the basic visual feature. At the same time, the visual features are extracted at 25Hz, since the audio features are processed at 100Hz, and the visual features are linearly interpolated to make them occur at the same frame rate as the audio features. Totally, the 60 dimensions lip geometrical features are obtained.

3.3 Experiment setup and results

To evaluate the performance of model proposed in the paper, for the sake of comparison, experiments are also done under the same conditions on the conventional triphone HMM, two single stream DBN model (Lv et al. 2007): WP-DBN models with word-phone structure and WPS-DBN model with word-phone-state structure, At the same time, state synchrony multi-stream HMM (SS-MSHMM) and state asynchrony multi-stream HMM (SA-MSHMM) are used, and SA-MSHMM is implemented as product HMM, with four audio and four video HMM states, totally 16 composite states. For PHMM, at various SNRs (ranging from 0dB to 30dB), stream exponent of audio stream is varied from 0 to 1 in step of 0.05, and the value of the stream exponent that maximized the word accuracy is chosen.

In the experiments on the digit audio-visual database, for MS-ADBN model, both in audio stream and in visual stream, each of the 22 phones is associated with the observation feature, with the probability is modeled by 1 Gaussian. Together with the three-phone "silence" model and the one phone "short pause" model which actually ties its phone with the middle phone of the silence model, totally there are 50 Gaussians in the model for two streams. While for MM-ADBN model, in each stream, each of the 22 phones is composed of 4 states modeled by 1 Gaussian, together with the silence and short pause model, totally, there are 182 Gaussian. Since in the training sentences, each digit has about 60 samples, and each phone has about 100 occurrences. It can be seen that every model can be properly trained.

In the experiments on the continuous audio-visual database, for MS-ADBN model, in each stream, each of the 74 phones is associated with the observation feature, with the probability is modeled by 1 Gaussian. Together with the silence and short pause model, totally parameters of 154 Gaussians need be trained for both streams. While for MM-ADBN model, in each stream, each of the 74 phones is composed of 4 states modeled by 1 Gaussian, together with silence model and short pause model, totally, there are 598 Gaussians. As a consequence, comparing with MM-ADBN model, MS-ADBN model has relatively small training parameters. In the training set, since each word has about 4 samples, MS-ADBN can not be trained sufficiently, while each phone has about 600 occurrences. MM-ADBN can be properly trained.

Setup	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
SS-MSHMM (audio and visual feature)	42.73	54.24	67.88	77.27	85.15	91.21	92.73	69.75
SA-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

Table 1. Word recognition rate for the digit audio-visual database (in %)

Setup	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
SA-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

Table 2. Word recognition rate for the continuous audio-visual database (in %)

Word recognition rates for digit audio-visual database and continuous audio-visual database, using MS-ADBN model and MM-ADBN model, respectively, are given in Table 1 and Table 2. For the sake of comparison, word recognition rates obtained from HMM, SS-MSHMM, SA-MSHMM, WP-DBN model and WPS-DBN model are also given.

It can be notice from Table 1 and Table 2 that:

a. For audio-only speech recognition on digit audio-visual database, under clean or relatively clean conditions with SNRs as 20dB and 30dB, the speech recognition rates of WP-DBN model are lower than those of triphone HMM. But the recognition rates under 20dB show that WP-DBN is more robust to noisy environments. Additionally, for speech recognition with visual features on digit audio-visual database, WP-DBN model performs slightly better than triphone HMM. A possible reason is that the DBN model describes better the dynamic temporal evolution of the speech process. While WPS-DBN model has the worse performance than triphone HMM, a possible reason is that

WPS-DBN model uses single Gaussian model, triphone HMM uses Multi-Gaussian mixture model. For audio only or video only speech recognition on continuous audio-visual database, WPS-DBN model outperform than triphone HMM at various SNRs.

b. Because of integrating the visual features and audio features, multi-stream models have the better performance than corresponding single stream models. For digit audio-visual database, in the noisy environment with signal to noise ratios ranging from 0dB to 30dB, comparing with HMM, WP-DBN and WPS-DBN model, the average improvements of 6.03%, 8.67% and 7.34% are obtained in speech recognition rate from SA-MSHMM, MS-ADBN and MM-ADBN model respectively. As well as for continuous audio-visual database, in clean speech, the improvements of 5.61%, 7.81% and 0.42% respectively.
c. For digit audio-visual database, MS-ADBN model has the better performance than SS-MSHMM and SA-MSHMM. This trend becomes even more obvious with the increasing of noise. Since the SA-MSHMM forces audio stream and visual stream to be synchronized at the timing boundaries of phones, while the MS-ADBN model looses the asynchrony of both streams to word level, the recognition results show the evidence that the MS-ADBN model describes more reasonable audio visual asynchrony in speech. As well as for continuous audio-visual database, MM-ADBN model has the better performance than SA-MSHMM. At clean speech environment, MM-ADBN model has the improvement of 9.97% than SA-MSHMM in speech recognition rate.
d. It should be noticed that under all noise conditions for digit audio-visual database, the MM-ADBN model gets worse but acceptable recognition rates than the MS-ADBN model, while for continuous audio-visual database, MM-ADBN model outperform than MS-ADBN model at various SNRs. At clean speech environment, the speech recognition rate of MS-ADBN model is 35.91% higher than that of the MS-ADBN in speech recognition rate. These are in coincidence with the speech recognition results of the single stream WP-DBN model and WPS-DBN model in (Lv et al. 2007). Since MM-ADBN model and WPS-DBN model are all phone models and are appropriate for large vocabulary speech recognition. MS-ADBN model and WP-DBN model are all word models, which cannot be properly trained for large vocabulary database, and they are appropriate for small vocabulary speech recognition, since they can be properly trained.

4. Conclusions and future work

In this paper, two multi-stream asynchrony Dynamic Bayesian Network (DBN) model: MS-ADBN model and MM-ADBN model, are proposed for small vocabulary and large vocabulary audio-visual speech recognition, which loose the limitation of asynchrony of the audio stream and visual stream to word level. Essentially, MS-ADBN model is a word model with word-phone-observation topology structure, whose recognition basic units are word, while MM-ADBN model is phone model with word-phone-state-observation topology structure, whose recognition basic units are phones. Speech recognition experiments are done on digit audio-vidio database and continuous audio-vidio database, results show that: MS-ADBN model has the highest recognition rate on digit audio-visual database; while for continuous audio-visual database, in clean speech environment, comparing with SA-MSHMM and MS-ADBN model, the improvements of 35.91% and 9.97% are obtained for MM-ADBN model in speech recognition rate. In the future work, we

will continue to improve the MM-ADBN model, and build up the MM-ADBN model based word-triphone-state topology for large vocabulary audio-visual speech recognition.

5. Acknowledgment

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 National High Technology Research and Development Program of China (Grant No. 2007AA01Z324). We would like to thank Prof. H. Sahli and W. Verhelst (Vrije Universiteit Brussel, Electronics & Informatics Dept., Belgium) for their help and for providing some guidance. We would also like to thank Dr. Ilse Ravyse, Dr. Jiang Xiaoyue, Dr. Hou Yunshu and Sun Ali for providing some help for the audio-visual database and visual feature data.

6. References

Lippmann, R. P. (1997). *Speech recognition by machines and humans*. speech communication, vol. 22, pp. 1-15, 1997.

Hermansky, H. (1990). *Perceptual Linear Predictive (PLP) Analysis of speech*. Journal of Acoustical Society of America, vol. 87, No. 4, pp. 1738-1752, 1990.

Hermansky, H. & Morgan, N. (1994). *RASTA processing of speech*. IEEE transaction on speech and audio processing, vol. 2, no.4, pp. 587-589, 1994.

Potamianos, G. & Neti, C. et al (2003). *Recent advances in the automatic recognition of audiovisual speech*. Proc. IEEE, vol.91, no 9, pp.1306-1326, 2003.

Dupont, S. & Luettin, J. (2000). *Audio-visual speech modeling for continuous speech recognition*. IEEE Trans. on Multimedia, vol. 2, pp.141–151, 2000.

Gravier, G.; Potamianos, G. & Neti, C. (2002). *Asynchrony modeling for audio-visual speech recognition*. in Proc. Human Language Technology Conf., San Diego, CA, pp. 1–6, 2002.

Nefian, A.; Liang, L. & Pi, L. et al (2002). *Dynamic Bayesian Networks for audio-visual speech recognition*. in EURASIP Journal on Applied Signal Processing, vol. 11, pp.1-15, 2002.

Bilmes, J. & Zweig, G. (2002). *The Graphical Modelds Toolkit:An Open Source Software System For Speech And Time-Series Processing*. Proceedings of the IEEE International Conf. on Acoustic Speech and Signal Processing (ICASSP), vol. 4, pp.3916-3919, 2002.

Murphy, K. (2002). *Dynamic Bayesian networks: Representation, inference and learning*. Ph.D. dissertation, Dept. EECS, CS Division, Univ. California, Berkeley, 2002.

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

Bilmes, J & Zweig, G. et al (2001). *Discriminatively structured graphical models for speech recognition: JHU-WS-2001 final workshop report*. Johns Hopkins Univ., Baltimore, MD, Tech. Rep. CLSP, 2001.

Lv, G.Y.; Jiang, D.M. & H, Sahli. et al (2007). *a Novel DBN Model for Large Vocabulary Continuous Speech Recognition and Phone Segmentation*. International Conference on Artificial Intelligence and Pattern Recognition (AIPR-07), Orlando, USA, pp. 437-440, 2007.

Zhang, Y.M.; Diao, Q. & Huang, S. et al (2003). *DBN based multi-stream models for speech*. in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Hong Kong, China, vol. 1, pp. 836–839, 2003.

Gowdy, J.N.; Subramanya, A. & Bartels, C. et al (2004). *DBN based multi-stream models for audio-visual speech recognition*. In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 1, pp. 993–996, 2004.

Bilmes, J. & Bartels, C. (2005). *Graphical Model Architectures for Speech Recognition*. IEEE Signal Processing Magazine, Vol. 22, no.5, pp.89–100, 2005.

Young, S.J.; Kershaw, D. & Odell, J. et al (1998). *The HTK Book*. http://htk.eng.cam.ac.uk/docs/docs.shtml.

Hirsch, H. G. & Pearce, D. (2000). *The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions*. ICSA ITRW ASR2000, September, 2000.

Zhou, Y.; Gu, L. & Zhang, H.J. (2003). *Bayesian Tangent Shape Model: Estimating Shape and Pose Parameters via Bayesian Inference*. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2003), Wisconsin, USA, vol. 1, pp. 109-116, 2003.

Ravyse, I. ; Jiang, D.M. & Jiang, X.Y. et al (2006). *DBN based Models for Audio-Visual Speech Analysis and Recognition*. PCM 2006, vol. 1, pp.19-30, 2006.

Speech Recognition

Edited by France Mihelic and Janez Zibert

ISBN 978-953-7619-29-9
Hard cover, 550 pages
Publisher InTech
Published online 01, November, 2008
Published in print edition November, 2008

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

INTECH
open science | open minds

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