Modeling the Synchrony between Audio and Visual Modalities for Speaker Identification

Yu Wang[1,3], Zhiyong Wu[2,3], Lianhong Cai[1,3] & Helen M. Meng[2,3]

1 Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
2 Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China,
3 Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
yuwang@mails.tsinghua.edu.cn, zywu@se.cuhk.edu.hk,
clh-dcs@tsinghua.edu.cn, hmmeng@se.cuhk.edu.hk

Abstract

This work aims to understand and model the inter-modal temporal relations between the audio and visual modalities of speech and validate whether the captured relations can improve the performance of audio-visual bimodal modeling for such applications as audio-visual speaker identification. We propose to extend our audio-visual correlative model (AVCM) with explicit durational modeling of the partial temporal synchrony between the two speech modalities, i.e. where the audio may lead, lag or remain synchronized with the video. We refer to the new extended model as Durational-AVCM. Experiments on the CMU database and a home-grown database demonstrate that Durational-AVCM can improve the accuracies of audio-visual speaker identification at all levels of acoustic signal-to-noise ratios (SNR) from 0dB to 30dB with varying acoustic conditions compared to original AVCM model. The results indicate the importance of incorporating the partial temporal synchrony between audio and visual modalities for audio-visual bimodal modeling.

1. Introduction

Human speech is produced by the movement of articulators. As some of these articulators are visible, there are inherent correlations between audio and visual speeches. There is also *partial temporal synchrony* between them. We may consider the process of human speech production, where the voice source originates from airflow from the lungs through laryngeal vibrations, producing quasi-periodic air pulses which are modulated by the vocal tract in different ways according to the positions of the articulators (e.g. lips, tongue, teeth and jaw). It is therefore conceivable that there should be dependence, or *direct synchrony*, between the visual signal which represents the movements of the articulators and the acoustic signal which represents the modulated airflow. Furthermore, it is possible for the articulators to move in anticipation of a phonetic event up to tens or hundreds of milliseconds prior to the actual production of the phone [1]. This means that the visual evidence will *temporally lead* the acoustic evidence. Conversely, due to co-articulatory effects and articulator inertia, the visual stream may *temporally lag* the acoustic evidence [2]. And sometimes, the two evidences might be *temporally synchronized*.

The audio-visual bimodal nature of speech has attracted significant interests in research community in recent years. It is believed there is much to be gained by leveraging the complementary and redundant relationships between audio and visual modalities to enhance human-computer speech communication, including audio-visual speech recognition, audio-visual speaker identification, etc [3-5].

Previous literatures generally divide the audio-visual integration strategies into three categories: feature-level fusion, decision-level fusion and model-level fusion [5-7]. In feature-level fusion, multiple features are concatenated into a large feature vector and a single model is trained [6] by assuming that the audio and visual features are *in strict synchrony*. In decision-level fusion, audio and visual features are processed separately to build two independent models [7], which assumes *complete independence* between the audio and visual features, and thus ignores the audio visual correlations. In model-level fusion, several models have been proposed, such as multi-stream hidden Markov model (HMM) [8], factorial HMM [8], coupled HMM [4], mixed dynamic Bayesian Networks (DBN) [9], etc. Multi-stream HMM and factorial HMM assume *independence* between audio and visual features. Coupled HMM and mixed DBN force audio visual streams to be *in strict synchrony* at model boundaries by introducing "anchor-points". We believe that model-level fusion is desirable as it offers flexibility in modeling *partial temporal synchrony*, for such applications as speaker identification, speech recognition, and speech synthesis etc.

Our previous work proposed an audio-visual correlative model (*AVCM*) [10] which is realized using the dynamic Bayesian networks (DBN), to describe both the inter-correlations and the loose temporal synchrony between audio and video streams. However, AVCM assumes that the temporal asynchrony between the audio and visual streams should be limited to at most *one state*. Moreover, the AVCM lacks the ability to model the audio-visual temporal relationship directly and explicitly. Hence the model has much room for improvement in modeling the inter-modal temporal relations between the audio and visual modalities.

The objective of the current paper is to propose the *Durational-AVCM* to capture the partial temporal synchrony. We reference durational modeling methods that have previously been used for HMMs. The approaches may be grouped into three categories. (i) Hidden semi-Markov models choose the occupancy of each state directly from a specified duration distribution. Examples include the explicit duration HMM [11] which learns a discrete duration distribution; as well as continuously variable duration HMM

[12] which learns a parametric duration distribution. (ii) Variable transition HMMs incorporate the transition probabilities of each state as a function of the state's current occupancy and allows for any duration distribution. Examples include inhomogenous HMM [13], nonstationary HMM [14], etc. (iii) Expanded state HMMs are standard HMMs with more states and/or state topologies, often coupled with state distribution tying [15], [16].

The outline of this paper is as follows: Section 2 describes the original AVCM. Section 3 presents the extension to Durational-AVCM. Experimental results and comparative analysis are given in section 4, which shows how Durational-AVCM outperforms original AVCM by modeling the partical temporal synchrony. Finally, section 5 concludes the paper.

2. DBN based audio-visual correlative model (AVCM)

Figure 1 illustrates our original audio-visual correlative model (AVCM) [10]. The figure shows a whole sentence model that consists of several words. The square nodes represent discrete variables, while the round nodes represent continuous variables. The hollow nodes represent hidden variables and the shaded nodes are observed. The upper part of the model describes the audio stream and the lower part describes the video stream. The labeled nodes include:

- Word Pos (WP): the current word index in the sentence;
- Word (W): the current word which is determined by the sentence and "Word Pos (WP)" node;
- State Pos (SP): the state index in the current word model;
- State (C): the current state which is determined by "Word (W)" and "State Pos (SP)" nodes;
- Observation (O): the audio / visual feature observations;
- EOS: represents the end of the sentence;
- Word Advance (WA): denotes the synchrony between audio and visual streams at word level. It may take the values 0, 1 or -1 to respectively denote that the two modalities are *synchronized*, that the audio *leads* the video, or that the audio *lags* the video;
- Last State (LS): indicates whether the current sate is the last state of the word;
- State Trans (T): can take the values *true* or *false* to indicate when the current state ends and switches to the next state;
- Word Trans (WT): may also take the values *true* or *false* to denote whether there is a word transition.

The "State Trans (T)" node depends on the "State" nodes from both streams, and also the "Word Advance" node. If the current stream is in leading, the "State Trans" node will always take the value *false* to prevent advancing more sates.

The "Word Trans (WT)" node depends on not only the "State Trans", "Last State", and "Word Advance" nodes of current stream, but also the "Last State" node of the other stream. Its value must be *false* in order to force the current stream to wait (despite cases where it may be read to advance to the next word) for the other stream to catch up (as long as the stream is not in its last sate).

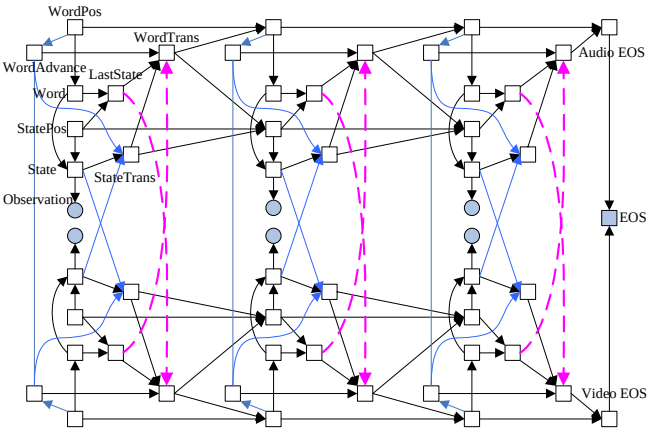

Figure 1: *Audio-Visual correlative model (AVCM).*

3. Durational-AVCM

As mentioned above, there are two defects of the original AVCM. First, the model topologies and node dependencies of "State Trans (T)" and "Word Trans (WT)" nodes limit the temporal asynchrony between two streams to within only *one state*. Second, the transition of words, i.e. "Word Trans", ignores the explicit temporal relationship between the two streams. We aim to incorporate explicit durational modeling with greater flexibility by proposing the *Durational-AVCM*, described in this section as follows.

3.1. Characteristics of the Durational-AVCM

To model the partial temporal synchrony between the audio and video streams (i.e. where the audio *leads*, *lags*, or is *synchronized* with video), we introduce the "Word Advance Duration (WD)" node to explicitly record and control the temporal relationship between the two streams, as depicted in figure 2.

For the advanced stream (Audio or Video), if the *WD* > *MaxAdvanceA* or *MaxAdvanceV (Word_A, Word_V)*, it should wait until the other stream catches up; otherwise, it could go on at its own step. Here, *MaxAdvanceA (Word_A, Word_V)* denotes the maximum advanced time counter when the audio stream is in advance while the current word of audio stream and video stream is *Word_A* and *Word_V* respectively. *MaxAdvanceV (Word_A, Word_V)* denotes the maximum value when video stream is in advance while the current word of audio stream and video stream is *Word_A* and *Word_V* respectively. The values of *MaxAdvanceA* and *MaxAdvanceV* are learnt from the training data.

In this way, the temporal relationship (i.e. the *lead* or *lag* time) between the audio and video stream at word level may be controlled within a reasonable range.

3.2. Word Advance Duration (WD)

The "Word Advance Duration (WD)" node depends on the "Word" nodes from audio and video streams, the "Word Advance" node in current time slice, as well as the "WD" node in the previous time slice.

The conditional probability distribution (CPD) of the "Word Advance Duration (WD)" node is defined as:

$$P\left(WD_t = d' \mid WD_{t-1} = d, WA_t = i, W_t^A = j, W_t^V = k\right)$$

$$= \begin{cases} 1 & if \quad d' = -1 & and \quad i = 0 & & \\ 1 & if \quad d' = MaxAdvanceA(j,k) & and \quad i = 1 & and \quad d = -1 & \\ 1 & if \quad d' = MaxAdvanceV(j,k) & and \quad i = -1 & and \quad d = -1 & \\ 1 & if \quad d' = d - 1 & and \quad i \neq 0 & and \quad d > 0 & \\ 1 & if \quad d' = 0 & and \quad i \neq 0 & and \quad d = 0 & \\ 0 & otherwise & & & \end{cases} \quad (1)$$

The "Word Advance Duration" variable actually serves as a time counter. Equation 1 illustrates the following:

- when the audio and video streams are in synchrony, WD keeps the value -1 (i.e. condition #1 in the equation);
- at the onset of asynchrony, WD is initialized to the maximum advanced time (i.e. condition #2 for audio in advance, or condition #3 for video in advance in the equation);
- WD decreases by 1 for each time slice (condition #4);
- WD stays unchanged when it reaches 0 (condition #5).

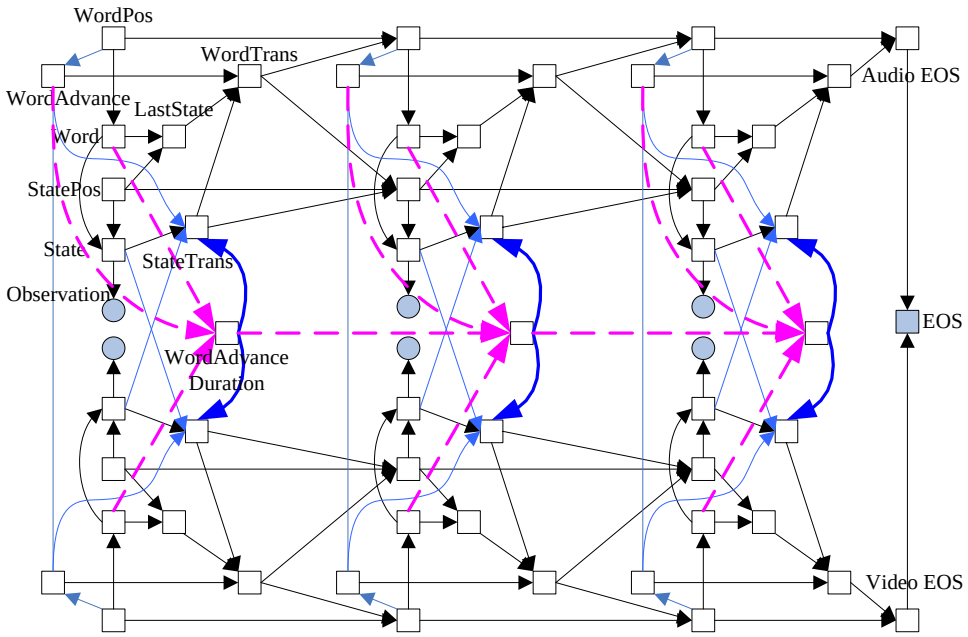

Figure 2: *The Durational-AVCM explicitly modeling the partial temporal synchrony between audio and visual streams by introducing the "Word Advance Duration (WD)" node.*

3.3. State Transition (T)

After introducing the "Word Advance Duration (WD)" node, the transition of the "State" in current word will depend on whether the asynchrony between the two streams has reached the maximum limitation. In the model, as depicted in figure 2, the "State Trans (T)" node depends on the "Word Advance Duration (WD)" node.

The CPD of the "State Trans (T)" node is defined as:

$$P\left(T_t^s = b \mid C_t^s = i, C_t^{\neg s} = j, WA_t = k, WD_t = d\right)$$

$$= \begin{cases} 1 & if \quad b = 0 \quad and \quad s = A \quad and \quad k = 1 \quad and \quad d = 0 \\ 1 & if \quad b = 0 \quad and \quad s = V \quad and \quad k = -1 \quad and \quad d = 0 \\ A_{ii,j}^s & if \quad b = 0 \quad and \quad s = A \quad and \quad k = 1 \quad and \quad d > 0 \\ 1 - A_{ii,j}^s & if \quad b = 1 \quad and \quad s = A \quad and \quad k = 1 \quad and \quad d > 0 \\ A_{ii,j}^s & if \quad b = 0 \quad and \quad s = V \quad and \quad k = -1 \quad and \quad d > 0 \\ 1 - A_{ii,j}^s & if \quad b = 1 \quad and \quad s = V \quad and \quad k = -1 \quad and \quad d > 0 \\ A_{ii,j}^s & if \quad b = 0 \quad and \quad s = A \quad and \quad k \neq 1 \\ 1 - A_{ii,j}^s & if \quad b = 1 \quad and \quad s = A \quad and \quad k \neq 1 \\ A_{ii,j}^s & if \quad b = 0 \quad and \quad s = V \quad and \quad k \neq -1 \\ 1 - A_{ii,j}^s & if \quad b = 1 \quad and \quad s = V \quad and \quad k \neq -1 \\ 0 & otherwise \end{cases} \quad (2)$$

Where:

- variable s represents the current stream;
- $\neg s$ represents the other stream (e.g. if s=A then $\neg s$=V, and vise versa);
- $A_{ii,j}^s$ is the probability for current stream s to stay at state i when the other stream is at state j. If the "WD" time counter exceeds the maximum advanced time (i.e. WD_t=d=0), the advanced stream should wait until the other stream catches up (condition #1 and #2 in the equation 2). Otherwise, the state of the stream just transits at its own pace according to the transition probability $A_{ii,j}^s$ (the other conditions in the equation 2).

As may be seen from equation (2), the *one state* limitation of AVCM for "State Trans (T)" node is relaxed in the Durational-AVCM.

3.4. Word Transition (WT)

As mentioned in Section 2, the "Word Trans (WT)" node of the AVCM depends on several factors. The Durational-AVCM relaxes the constraints and the transition of "Word" in one stream depends only on the "State Trans (ST)" of the same stream. However, it should be noted that the "Sate Trans (ST)" takes into account the word asynchrony between two streams, through the use of the "Word Advance Duration (WD)" node, as described in equation (2). Hence, the Durational-AVCM can explicitly model the temporal

asynchrony *within one word* between the two streams for different word combinations.

The CPD of the "Word Trans (WT)" node is defined as:

$$P\left(WT_t^s = b \mid LS_t^s = f, T_t^s = g, WA_t = i\right) = \begin{cases} 1 \;\; if \;\; b=0 \;\; and \;\; f=0 \\ 1 \;\; if \;\; b=0 \;\; and \;\; f=1 \;\; and \;\; g=0 \\ 1 \;\; if \;\; b=1 \;\; and \;\; f=1 \;\; and \;\; g=1 \;\; and \;\; i=-1 \; and \; s=A \\ 1 \;\; if \;\; b=1 \;\; and \;\; f=1 \;\; and \;\; g=1 \;\; and \;\; i=1 \;\; and \; s=V \\ 1 \;\; if \;\; b=1 \;\; and \;\; f=1 \;\; and \;\; g=1 \;\; and \;\; i=0 \\ 0 \;\; otherwise \end{cases} \quad (3)$$

4. Experiments

In order to validate whether modeling the partial temporal synchrony between audio and visual streams can really improve the performance of audio-visual bimodal model, we perform the following experiments to compare the performance of Coupled-HMM (CHMM), original AVCM and the new proposed Durational-AVCM.

Because we focus on validating the benefits of explicitly modeling the partial temporal synchrony for audio-visual bimodal modeling, relatively simple experiments on text-prompts speaker identification are performed as the test bed. The speaker identification results from CHMM, the original AVCM and Durational-AVCM are compared.

The experiments are based on the audio-visual bimodal database from Carnegie Mellon University (CMU database) [17] as well as our own home-grown database. The CMU database includes 10 subjects (7 males and 3 females) speaking 78 isolated words repeated 10 times. These words include numbers, weekdays, months, and others that are commonly used for scheduling applications. The home-grown database includes 60 subjects (38 males and 22 females, aged from 20 to 65) with each subject speaks 30 connect-digit words (the digit length differs from 2 to 6), and each utterance is repeated three times at intervals of one month.

The acoustic front-end includes 13 Mel frequency cepstral coefficients (MFCCs) and 1 energy parameter (with frame window size of 25ms and frame shift of 11ms) together with their delta parameters. Hence the audio feature vector has 28 dimensions.

The visual front-end includes mouth width, upper lip height, lower lip height [17] and their delta parameters. Thus the visual feature vector has 6 dimensions. The video frame rate is 30 frames per second (fps), which is up-sampled to 90fps (11ms) by copying and inserting two frames between each two original video frames. It should be noted that the visual features might be a bit too simple for speaker identification (as more complicate facial features can be used instead), however, given the purpose of the experiment to validate the benefits of modeling the partial temporal synchrony, it is enough already.

Artificial white Gaussian noise was added to the original audio data (SNR=30dB) to simulate various SNR levels. The models were trained at 30dB SNR and tested under SNR levels ranging from 0dB to 30dB at 10dB intervals. We applied cross-validation for every subject's data, i.e. 90% of all the data are used as training set, and the remaining 10% as testing set. The partitioning was repeated until all the data had been covered in the testing set.

All the models are implemented as DBNs. A DBN is developed for each word, with a left-to-right no skipping logical structure, which means the no. of the new transited state is always equal to or 1 greater than the original state. The audio model has 5 states, and video model has 3 states, each state is modeled using the Gaussian mixture model (GMM) with 3 mixtures. During speaker identification, the words' DBNs are connected to form a whole sentence model, which is then used to identify the speakers. The DBNs are implemented using the GMTK toolkit [18].

Table 1: *Average accuracies (%) of speaker identification under different SNR on CMU database.*

audio signal-to-noise ratio (SNR)	30dB	20dB	10dB	0dB
CHMM	100	88	79	60
Original AVCM	100	92	79	65
Durational-AVCM	100	93.5	82	69

Table 2: *Average accuracies (%) of speaker identification under different SNR on own homegrown database.*

audio signal-to-noise ratio (SNR)	30dB	20dB	10dB	0dB
CHMM	100	85	77	57
Original AVCM	100	89	78	61
Durational-AVCM	100	91	80	66

The identification accuracies from all the testing data are averaged and reported as the final result. The results on CMU database are summarized in table 1. The experiments are also conducted on our own homegrown database with a larger number of speakers and results are summarized in table 2.

As can be seen from the results that the original AVCM model outperforms Coupled-HMM (CHMM) as it can, to some extend, describe the temporal synchrony between audio and video stream *in one state*. Moreover, the Durational-AVCM model proposed in this paper has even higher accuracy than original AVCM. This indicates that the Durational-AVCM can model the *partial temporal synchrony* even with more accuracy which indicates that the new proposed Durational-AVCM is efficient. The results also indicate that it is important to model the partial temporal relationships (partial synchrony) between audio and video streams in the audio-visual bimodal modeling.

5. Conclusions

This paper investigates the partial temporal synchrony between audio and video streams. A new durational audio-visual correlative model (*Durational-AVCM*) is proposed, which explicitly models the maximum temporal asynchrony (i.e. where the audio *leads* video, audio *lags* video, and/or audio is *synchronized* with video) between the audio and video streams at the *word level*. The experiments on the audio-visual bimodal speaker identification demonstrate that the Durational-AVCM model improves the identification accuracies compared to the coupled HMM and original AVCM. The results also indicate that it is important to model temporal synchrony between audio and visual modalities for audio-visual bimodal applications such as speaker identification.

Our further work will focus on this to explicitly model the state durations at the *state level*. We will also investigate the use of Durational-AVCM for audio-visual speech recognition (AVSR) and text-to-audio-visual-speech synthesis (TTAVS).

6. Acknowledgments

This work is jointly supported by the research fund from the National Natural Science Foundation of China (NSFC) under grant No. 60433030, 60418012 and the Hong Kong SAR Government's Research Grants Council (RGC) Earmarked Grant No. CUHK4149/06E.

7. References

[1] Benoit, C., 1992. The Intrinsic Bimodality of Speech Communication and the Synthesis of Talking Faces. *Journal on Communications of the Scientific Soc. For Telecommunications*, vol. 43, 32-40.

[2] Hazen, T., 2006. Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition. *IEEE Trans. on Speech & Audio Proc.*, vol. 14(3), 1082-1089.

[3] Senior, A.; Neti, C.; Maison, B., 1999. On the Use of Visual Information for Improving Audio-based Speaker Recognition. *Proc. Audio-visual Speech Processing Conf.*, 108-111.

[4] Nefian, A.V.; Liang, L.H.; Fu, T.Y.; Liu, X.X., 1999. A Bayesian Approach to Audio-Visual Speaker Identification. *Proc. 4th International Conf. Audio- and Video-based Biometric Person Authentication*, vol. 2688, 761-759.

[5] Chibelushi, C.C.; Deravi, F.; Mason, J.S.D., 2002. A Review of Speech-based Bimodal Recognition. *IEEE Trans. Multimedia*, vol. 4, 23-37.

[6] Chibelushi, C.C.; Mason, J.S.D.; Deravi, F., 1997. Feature-level Data Fusion for Bimodal Person Recognition. *Proc. 6th Int. Conf. Image Processing and its Application*. IEEE, 399-403.

[7] Chatzis, V.; Bors, A.G.; Pitas, I., 1999. Multimodal Decision-level Fusion for Person Authentication. *IEEE Trans. Syst. Man Cybern. A*, vol. 29, 674-680.

[8] Depont, S.; Luettin, J., 2000. Audio-Visual Speech Modeling for Continuous Speech Recognition. *IEEE Trans. Multimedia*, vol. 2(3), 141-151.

[9] Gowdy, J.N.; Subramanya, A.; Bartels, C.; Bilmes, J., 2004. DBN based Multi-stream Models for Audio-Visual Speech Recognition. *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*. IEEE, vol. 1, 993-996.

[10] Wu, Z.Y.; Cai, L.H.; Meng, H.M., 2006. Multi-level Fusion of Audio and Visual Features for Speaker Identification. *Proc. Int. Conf. Biometrics*, 493-499.

[11] Ferguson, J.D., 1980. Variable Duration Models for Speech. *Proc. Symp. App. Hidden Markov Models Text Speech.*

[12] Levinson, S.E., 1986. Continuously Variable Duration Hidden Markov Models for Speech Analysis. *Proc. Int. Conf. Acoust., Speech, Signal Processing.* 1241-1244.

[13] Ramesh, P.; Wilpon, J.G., 1992. Modeling State Durations in Hidden Markov Models for Automatic Speech Recognition. *Proc. Int. Conf. Acoust., Speech, Signal Processing.* 381-384.

[14] Sin, B.; Kim, J.H., 1995. Nonstationary Hidden Markov Model. *IEEE Trans. Signal Processing*, vol. 46, 31-46.

[15] Russell, M.J.; Cook, A.E., 1987. Experimental Evaluation of Duration Modeling Techniques for Automatic Speech Recognition. *Proc. Int. Conf. Acoust., Speech, Signal Processing*. 2376-2379.

[16] Russell, M.J.; Moore, R.K., 1985. Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition. *Proc. Int. Conf. Acoust., Speech, Signal Processing*. 5-8.

[17] Chen, T., 2001. Audiovisual Speech Processing. *IEEE Trans. Signal Processing*, vol. 18, 9-21.

[18] Bilmes, J.; Zweig, G., 2002. The Graphical Models Tookit: An Open Source Software Systems for Speech and Time-series Processing. *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, IEEE, vol. 4, 3916-3919.