15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland, September 3-7, 2007, copyright by EURASIP

MULTIMODAL FUSION FOR CUED SPEECH LANGUAGE RECOGNITION

Savvas Argyropoulos[1,2]*, Dimitrios Tzovaras*[2]*, and Michael G. Strintzis*[1,2]

[1] Electrical and Computer Engineering Dept.
Aristotle University of Thessaloniki
Thessaloniki, GR-54124, Hellas

[2] Informatics and Telematics Institute
Centre for Research and Technology Hellas
Thermi-Thessaloniki, GR-57001, Hellas

e-mail: {savvas@ieee.org, Dimitrios.Tzovaras@iti.gr, strintzi@eng.auth.gr}

ABSTRACT

In this paper, a novel method for Cued Speech language recognition is proposed. A multimodal processing framework is developed for the efficient fusion of the modalities under consideration. The robust feature extraction from the audio, lip shape and gesture modalities is also discussed. Moreover, the inherent correlation among the signals is modelled using a modified Coupled Hidden Markov Model. The contribution of each modality to the fused decision is modified by assessing its reliability and assigning a weighting factor to improve the accuracy of the recognized words. The method is experimentally evaluated and is shown to outperform methods that rely only on the perceivable information to infer the transmitted messages.

1. INTRODUCTION

Recent technological advances have improved communication between disabled people. The emerging artificial intelligence techniques are starting to diminish the barriers for impaired people and change the way individuals with disabilities communicate. A common problem in intercommunication between impaired individuals is that, in general, they do not have access to the same modalities and the perceived communicated message is limited by one's disabilities. A quite challenging task involves the transformation of a signal into another perceivable form to enable or enhance communication. Ideally, a recognition system should combine all incoming modalities of an individual, perform recognition of the transmitted message, and translate it into signals that are more easily understood by impaired individuals.

Automatic translation of gestural languages has been an active topic during the last years as it is expected to improve everyday life of impaired people [1]. Much effort has been concentrated on the automatic recognition and generation of Cued Speech language [2], which is a specific gestural language (different from the sign language) used for communication between deaf and hearing people. An automatic cue generating system based on speech recognition is presented in [3], while a Cued Speech synthesizer was developed in [4] based on 3D hand and face movements. Additionally, a Cued Speech hand gesture recognition tool is presented in [5]. A glove is employed to improve hand shape segmentation and a set of parameters is used to construct a structural model. In [6], the limitation of using a glove is suppressed by a technique for automatic gesture recognition based on 2D and 3D methods.

This work was supported by the EC IST-NoE FP6-507609 (SIMILAR).

An automatic Cued Speech recognition system should fuse the input of three modalities, (audio, lip shape, and hand shape) and take advantage of their complementary nature to deduce correctly the transmitted message. Multimodal approaches have been shown to be advantageous in continuous audio-visual speech processing. In [7], audio and visual features are integrated using a Coupled Hidden Markov Model (CHMM). Moreover, the complementary and supplementary information conveyed by speech and lip shape modalities is investigated in [8], where CHHMs are also employed to exploit the interdependencies between these two modalities. CHMMs are also employed in [9] to model loosely-coupled modalities where only the onset of events is coupled in time. Moreover, the use of Dynamic Bayesian Networks (DBNs) is introduced to fuse the feature vectors extracted from lip shapes and the audio signal in [10].

In this paper, a multimodal fusion method for Cued Speech language recognition is proposed. The basic idea is to exploit the correlation among modalities to enhance the perceivable information by an impaired individual who can not access all incoming modalities. Thus, a modality which would not be perceived due to a specific disability can be employed to improve the information that is conveyed in the perceivable modalities and increase the accuracy rates of recognition. In that way, modality replacement can be achieved.

Based on that, a new multimodal fusion framework for continuous Cued Speech language recognition is proposed. A specific feature extraction method is presented for each modality involved in Cued Speech. Furthermore, the complex interactions and inter-dependencies among the modalities are modelled by a modified CHMM, in which the contribution of each modality is adjusted by a reliability measure. The common way of incorporating these reliability values into decision fusion is to use them as weighting coefficients and to compute a weighted average [11]. Modality reliability has also been examined in [12], in the context of multimodal speaker identification. The main contribution of this paper to the Cued Speech language recognition is the development of a fusion scheme for word recognition. To the best of our knowledge, this is the first approach to combine audio, lip shape, and hand shapes for the improvement of the recognition of the transmitted message in the context of Cued Speech language.

2. CUED SPEECH

The architecture of the proposed modality replacement approach is depicted in Fig. 1. The performance of such a system is directly dependent on the efficient multimodal processing of multiple inputs and the effective exploitation of

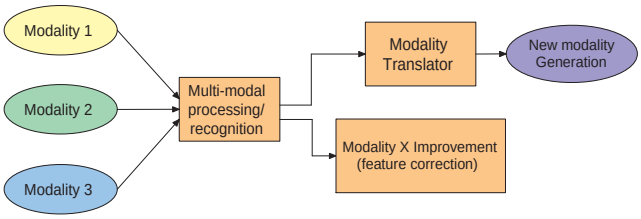

Figure 1: The modality replacement system architecture.

their mutual information to achieve accurate recognition of the transmitted content. After the recognition is performed effectively, either a modality translator can be employed to generate a new modality or the output can be utilized to detect and correct possibly erroneous feature vectors that may correspond to different modalities.

The Cued Speech paradigm has been selected to demonstrate the modality replacement concept. It uses eight hand shapes placed at four different positions near the face, as illustrated in Fig. 2, in combination with the natural lip movements of speech to make the sounds of spoken language look different from each other. In that sense, it can be considered as the visible counterpart of the spoken language, in which one uses visual information extracted from the speaker's lips to improve word recognition, especially in noisy environments. While talking, cued speakers execute a series of hand and finger gestures near the face closely related to what they are pronouncing. Thus, the transmitted message is contained into three modalities: audio, lip shapes, and hand shapes.

Cued Speech is based on a syllabic decomposition: the message is formatted into a list of "Consonant-Vowel syllables", which is known as CV list. Each CV is coded using a different hand shape, which is combined with the lip shape, so that it is unique and understandable.

The fact that hand shapes are made near the face and also that the exact number and orientation of fingers has to be determined in order to deduce the correct gesture differentiate Cued Speech from sign languages. In sign languages, the segmentation of the head and the hands is based on a skin color technique [13]. However, vision-based extraction of the hand features is prone to introducing errors when background color is similar to the skin color. Moreover, the skin color depends on the user and it is affected by lighting, background and reflected light from users clothes.

Since the scope of the present work lies mainly in the efficient combination and fusion of the modalities, the feature extraction and representation procedure is constrained to a brief analysis, given in the following section.

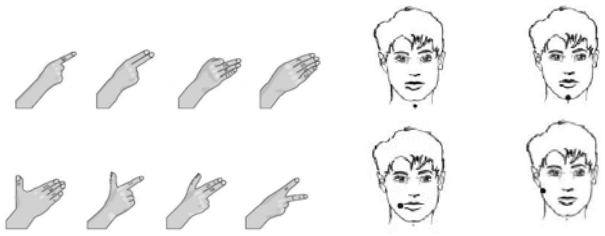

Figure 2: Hand shapes and face positions in Cued Speech Language.

3. FEATURE EXTRACTION AND REPRESENTATION

3.1 Audio feature extraction

Robust features have to be extracted from the audio signal to ensure acceptable recognition rates under various environments. The most popular features used for speech processing are the Mel Frequency Cepstral Coefficients (MFCC) since they yield good discrimination of the speech signal. The audio stream is processed over 15 *msec* frames centered on 25 *msec* Hamming window. The MFCC feature vector for a frame comprises 13 MFCCs along with their first and second order derivatives.

3.2 Lip shape feature extraction

For the lip shape modality, the robust location of facial features and especially the location of the mouth region is crucial. Then, a discriminant set of visual observation vectors have to be extracted. The process for the extraction of the lip shape is presented in [10], and is described in brief below so that the paper is self-contained.

Initially, the speaker's face is located in the video sequence as illustrated in Fig. 3. Subsequently, the lower half of the detected face is selected as an initial candidate of the mouth region and Linear Discriminant Analysis (LDA) is used to classify pixels into two classes: face and lip. After the lip region segmentation has been performed the contour of the lips is obtained using the binary chain encoding method and a normalized 64x64 region is obtained from the mouth region using an affine transform. In the following, this area is split into blocks and the 2D-DCT transform is applied to each of these blocks and the lower frequency coefficients are selected from each block, forming a vector of 32 coefficients. Finally, LDA is applied to the resulting vectors, where the classes correspond to the words considered in the application. A set of 15 coefficients, corresponding to the most significant generalized eigenvalues of the LDA decomposition is used as the lip shape observation vector.

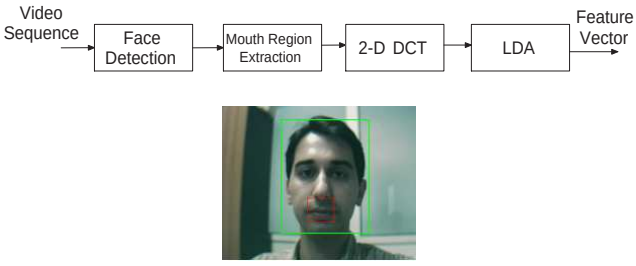

Figure 3: Mouth location for lip shape representation.

3.3 Gesture feature extraction

Since Cued Speech is strongly dependent on the hand shapes made by the cue speaker while talking, it is essential for a multimodal speech recognition system to detect and classify gestures in a correct and efficient manner. The main problem that has to be addressed is the segmentation of the hand and

the fingers in front of the cue speaker's face and the possible occlusions, which pose limitations to the application of a skin color mask. Moreover, since the exact orientation and the number of fingers is crucial the algorithm must be able to discriminate between fingers and estimate the position of the hand shape relative to the speaker's mouth.

To overcome the aforementioned problems the proposed method is based on gestures made with a glove consisting of 6 colours: one for each finger and one for the palm. In that way, a color mask can be employed to determine which finger is present in each frame and estimate the position of that finger. The mass centre of each finger is computed and the absolute difference between the mass centre of the mouth (as computed in the lip shape extraction procedure) and the mass centre of each finger forms the feature vector for the hand shape modality. It must be stressed that other parameters could also be selected to represent the hand gestures, such as the parameters employed in [5].

4. MULTIMODAL FUSION

The combination of multiple modalities for inference has proven to be a very powerful way to increase detection and recognition performance. By combining information provided by different models of the modalities, weakly incorrect evidence in one modality can be corrected by another modality. Hidden Markov Models (HMMs) are a popular probabilistic framework for modelling processes that have structure in time. Especially, for the applications that integrate two or more streams of data, Coupled Hidden Markov Models (CHMMs) have been developed.

A CHMM can be considered as a collection of HMMs, one for each data stream, where the hidden backbone nodes at time t for each HMM are conditioned by the backbone nodes at time $t-1$ for all the related HMMs. It must be noted that CHMMs are very popular among the audio-visual speech recognition community, since they can model efficiently the endogenous asynchrony between the speech and lip shape modalities. The parameters of a CHMM are described below:

$$\pi_0^c(i) = P(q_t^c = i) \quad (1)$$

$$b_t^c(i) = P(\mathbf{O}_t^c | q_t^c = i) \quad (2)$$

$$a_{i|j,k,n}^c = P(q_t^c = i | q_{t-1}^A = j, q_{t-1}^L = k, q_{t-1}^G = n) \quad (3)$$

where q_t^c is the state of the coupled node in the $c-th$ stream at time t, $\pi_0^c(i)$ is the initial state probability distribution for state i in the $c-th$ stream, $\mathbf{O}_t^c$ is the observation of the nodes at time t in the $c-th$ stream, $b_t^c(i)$ is the probability of the observation given the i state of the hidden nodes in the $c-th$ stream, and $a_{i|j,k,n}^c$ is the state transitional probability to node i in the $c-th$ stream, given the state of the nodes at time $t-1$ for all the streams. The distribution of the observation probability is usually defined as a continuous Gaussian mixture. Fig. 4 illustrates the CHMM employed in this work. Square nodes represent the observable nodes whereas circle nodes denote the hidden (backbone) nodes.

One of the most challenging tasks in automatic speech recognition systems is to increase robustness to environmental conditions. Although the stream weights needs to be properly estimated according to noise conditions, they can not be determined based on the maximum likelihood criterion.

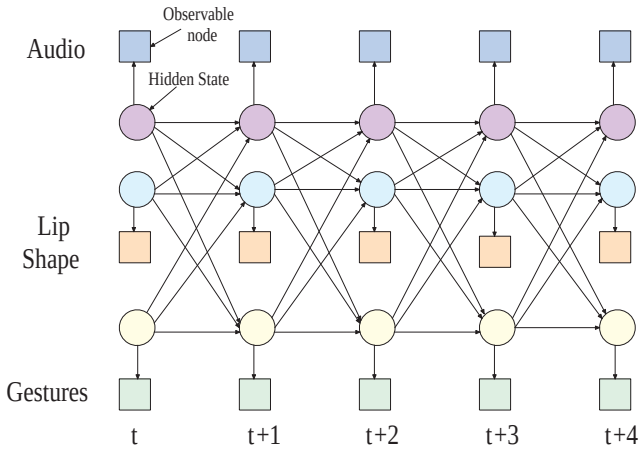

Figure 4: Coupled HMM for fusion of three modalities.

Therefore, it is very important to build an efficient stream-weight optimization technique to achieve high recognition accuracy.

4.1 Modality reliability

Ideally, the contribution of each modality to the overall output of the recognition system should be weighted according to a reliability measure. This measure denotes how each observation stream should be modified and acts as a weighting factor. In general, it is related to the environmental conditions (*e.g.*, acoustic noise for the speech signal).

The common way of incorporating these reliability values into decision fusion is to use them as weighting coefficients and to compute a weighted average. Thus, the probability $b_m(\mathbf{O}_t)$ of a feature $\mathbf{O}_t$ for a word m is given by:

$$b_m(\mathbf{O}_t) = w_A \cdot b_A(\mathbf{O}_t^A) + w_L \cdot b_L(\mathbf{O}_t^L) + w_G \cdot b_G(\mathbf{O}_t^G) \quad (4)$$

where $b_A(\mathbf{O}_t^A)$, $b_L(\mathbf{O}_t^L)$, $b_G(\mathbf{O}_t^G)$, are the likelihoods for an audio feature $\mathbf{O}_t^A$, a lip shape feature $\mathbf{O}_t^L$, and a gesture feature $\mathbf{O}_t^G$, respectively. The parameters w_A, w_L, and w_G are the audio, lip shape, and gesture weights, respectively, and $w_A + w_L + w_G = 1$.

In the proposed method, a different approach is employed to determine the weights of each data stream. More specifically, for each modality, word recognition is performed using a HMM for the training sequences. The results of the (unimodal) word recognition indicate the noise levels in each modality and provide an approximation of their reliability. More specifically, when the unimodal HMM classifier fails to identify the transmitted words it means that the observation features for the specific modality are unreliable. On the other hand, a small word error rate using only one modality and the related HMM means that the corresponding feature vector is reliable and should be favoured in the CHMM.

4.2 Modified Coupled Hidden Markov Model

4.2.1 Training

The Maximum Likelihood (ML) training of the dynamic Bayesian Networks in general and of the CHMMs in particular, is a well understood technique [10]. However, the

iterative maximum likelihood estimation of the parameters only converges to a local optimum, making the choice of the initial parameters of the model a critical issue. In this section, an efficient method for the initialization of the ML training that uses a Viterbi algorithm [14] derived for the CHMM is presented. The Viterbi algorithm determines the optimal sequence of states for the coupled nodes of audio, lip shapes, and hand shapes that maximize the observation likelihood. The Viterbi algorithm for the three-stream CHMM used in the developed system can be defined in four stages: initialization, recursion, termination, and backtracking.

1) Initialization

$$\delta_0(i,j,\vartheta) = \pi_0^A(i)\pi_0^L(j)\pi_0^G(\vartheta)b_t^A(i)b_t^L(j)b_t^G(\vartheta) \quad (5)$$

$$\psi_0(i,j,\vartheta) = 0 \quad (6)$$

2) Recursion

$$\delta_t(i,j,\vartheta) = \max_{k,l,m}\{\delta_{t-1}(k,l,m)a_{i|k,l,m}a_{j|k,l,m}a_{\vartheta|k,l,m}\}\cdot \cdot b_t^A(k)b_t^L(l)b_t^G(m) \quad (7)$$

$$\psi_t(i,j,\vartheta) = \arg\max_{k,l,m}\{\delta_{t-1}(k,l,m)a_{i|k,l,m}a_{j|k,l,m}a_{\vartheta|k,l,m}\} \quad (8)$$

3) Termination

$$P = \max_{i,j,\vartheta}\{\delta_T(i,j,\vartheta)\} \quad (9)$$

$$\{q_T^A, q_T^L, q_T^G\} = \arg\max_{i,j,\vartheta}\{\delta_T(i,j,\vartheta)\} \quad (10)$$

4) Backtracking

$$\{q_t^A, q_t^L, q_t^G\} = \psi_{t+1}(q_{t+1}^A, q_{t+1}^L, q_{t+1}^G) \quad (11)$$

4.2.2 *Recognition*

The word recognition is performed using the Viterbi algorithm, described above, for the parameters of all the word models. It must be emphasized that the influence of each stream is weighted at the recognition process because, in general, the reliability and the information conveyed by each modality is different. Thus, the observation probabilities are modified as:

$$b_t^A(i) = b_t(\mathbf{O}_t^A|q_t^A = i)^{w_A} \quad (12)$$

$$b_t^L(i) = b_t(\mathbf{O}_t^L|q_t^L = i)^{w_L} \quad (13)$$

$$b_t^G(i) = b_t(\mathbf{O}_t^G|q_t^G = i)^{w_G} \quad (14)$$

where w_A, w_L, and w_G are respectively the weights for audio, lip shape, and gesture modalities and $w_A + w_L + w_G = 1$. The values of w_A, w_L, and w_G are obtained using the methodology of section 4.1.

5. EXPERIMENTAL RESULTS

Since there is no available database for Cued Speech language, the multimodal speech recognition system was evaluated on multimodal sequences including gesture, lip and speech data corresponding to a small set of words. More specifically, the vocabulary consisted of digits from zero to nine. Each word of the considered vocabulary was repeated twelve times; nine instances of each word were used for training and three instances were used for testing. Five subjects were trained to perform the specified gestures.

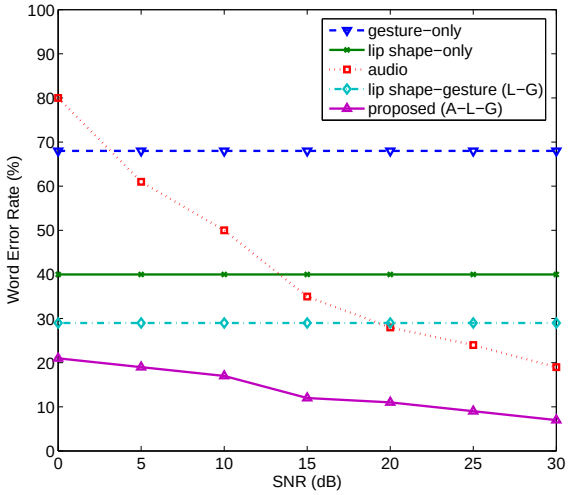

Figure 5: Word error rate at various acoustic noise levels

Six states were considered for the coupled nodes in the CHMM, with no back transitions and three mixture per state. The form of the feature vectors has been described in Section 3. For audio only, lip shape only, and gesture only recognition, a HMM was used with six states and no back transitions. In the audio-only and all multimodal Cued Speech recognition experiments including audio, the audio sequences used in training were captured in clean acoustic conditions. The whole system was evaluated for different levels of noise, in terms of Signal to Noise Ratio (SNR), and the results are depicted in Fig. 5.

The experimental results indicate that the proposed approach which exploits the audio, lip shape and gesture modalities (A-L-G) achieves smaller word error rate compared to the unimodal approaches. More specifically, the proposed system achieves consistently better results and the results are even more impressive compared to the audio-only method when the quality of the noise signal deteriorates (small SNR values). A reduction of approximately 20% in word error rate is achieved at 15 *dB*.

To demonstrate the merit of the proposed approach, a comparison with a scheme which exploits only visual information (lip shapes and hand shapes) is presented. This scenario corresponds to an individual with hearing disabilities while trying to understand a cue speaker. Due to hearing loss, the impaired person can rely only on lip and hand shapes to infer the transmitted message. However, much of the transmitted information is contained in the audio signal which the disabled person can not perceive. The results illustrated in Fig. 5 indicate that the Lip-Gesture (L-G) approach yields substantially lower recognition rates compared to the scheme which combines audio, lip shapes, and gestures (A-L-G).

The superior performance of the proposed system can be attributed to the effective modelling of the interactions

and the inter-dependencies among the three modalities. It is worth noting that the algorithm can correctly identify words even if all unimodal recognizers fail to recognize the transmitted word. As depicted in Table 1, the three unimodal recognizers misinterpret the word "four". However, the multimodal scheme can efficiently process the feature vectors from each modality and infer the correct word.

Table 1: Word recognition at 30 *dB* using different fusion methods.

Method	zero	one	two	three	four	five	six	seven	eight	nine
A-L-G	zero	one	two	three	four	five	six	eight	eight	nine
L-G	zero	zero	two	three	five	five	six	eight	eight	nine
Audio	zero	one	five	three	five	five	six	eight	eight	nine
Lip	zero	zero	two	three	five	five	six	eight	eight	one
Gesture	one	zero	two	three	five	five	zero	one	eight	nine

Future work will focus on implementing the modality replacement concept as depicted in an instance of the recognition procedure in Fig. 6. In particular, after the recognition of the transmitted message, a modality translation method should be applied to transform the conveyed information into a modality that can be perceived by the listener (e.g., text, haptic, audio, sign language, *etc*.) and, thus, improve inter-communication between disabled people. Moreover, although the proposed system was not tested in large vocabulary continuous Cued Speech recognition, it is expected that the proposed method can be extended without major modifications.

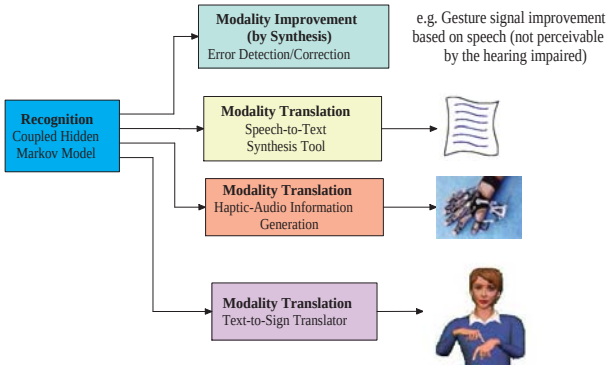

Figure 6: Potential applications of the modality replacement framework.

6. CONCLUSIONS

A novel multimodal fusion framework for continuous Cued Speech language recognition was proposed. A specific feature extraction method for the speech, lip shape, and gesture signals was presented for the efficient representation of the modalities. Furthermore, a modified CHMM was developed to fuse the multiple modalities and a reliability measure was estimated and applied on the data streams. The experimental evaluation demonstrated that by exploiting the correlation among all incoming modalities the recognition rate can be improved.

REFERENCES

[1] B. Bauer, S. Nießen, and H. Hienz, "Towards an Automatic Sign Language Translation System," in *Proc. of the International Workshop on Physicality and Tangibility in Interaction, Siena, Italy*, 1999.

[2] R. O. Cornett, "Cued speech," *American Annals for the Deaf*, vol. 112, pp. 3–13, 1967.

[3] P. Duchnowski, D.S. Lum, J.C. Krause, M.G. Sexton, M.S. Bratakos, and L.D. Braida, "Development of Speechreading Supplements Based on Automatic Speech Recognition," *IEEE Trans. on Biomedical Engineering*, vol. 47, no. 4, pp. 487–496, 2000.

[4] G. Gibert, G. Bailly, F. Elisei, D. Beautemps, and R. Brun, "Evaluation of a speech cuer: from motion capture to a concatenative text-to-cued speech system," in *Language Resources and Evaluation Conference (LREC), Lisbon, Portugal*, 2004.

[5] T. Burger, A. Caplier, and S. Mancini, "Cued speech hand gestures recognition tool," in *Proc. of the European Signal Processing Conference*, Antalya, Turkey, 2005.

[6] A. Caplier, L. Bonnaud, S. Malassiotis, and M.G. Strintzis, "Comparison of 2d and 3d analysis for automated cued speech gesture recognition," in *SPECOM*, 2004.

[7] H. Yashwanth, H. Mahendrakar, and S. David, "Automatic speech recognition using audio visual cues," in *Proc. of the First India Annual Conf., India*, 2004.

[8] X. Liu, Y. Zhao, X. Pi, L. Liang, and A.V. Nefian, "Audio-visual continuous speech recognition using a coupled hidden Markov model," in *Proc. of the Int. Conference on Spoken Language Processing (ICSLP)*, 2002.

[9] TT Kristjansson, BJ Frey, and TS Huang, "Event-coupled hidden Markov models," in *IEEE Int. Conf. on Multimedia and Expo*, 2000.

[10] A.V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1274–1288, 2002.

[11] S. Tamura, K. Iwano, and S. Furui, "A Stream-Weight Optimization Method for Multi-Stream HMMS Based on Likelihood Value Normalization," 2005.

[12] E. Erzin, Y. Yemez, and AM Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 840–852, 2005.

[13] N. Tanibata, N. Shimada, and Y. Shirai, "Extraction of hand features for recognition of sign language words," in *The 15th International Conf. on Vision Interface*, 2002.

[14] LR Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.