# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| Title | A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs |
| Author | Satoshi Tamura, Koji iwano, Sadaoki Furui |
| Journal/Book name | IEEE ICASSP 2004, Vol. , No. 1, pp. 857-860 |
| 発行日 / Issue date | 2004, 5 |
| 権利情報 / Copyright | (c)2004 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE. |

# A STREAM-WEIGHT OPTIMIZATION METHOD FOR AUDIO-VISUAL SPEECH RECOGNITION USING MULTI-STREAM HMMS

*Satoshi Tamura, Koji Iwano and Sadaoki Furui*

Tokyo Institute of Technology
Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
{tamura,iwano,furui}@furui.cs.titech.ac.jp

## ABSTRACT

For multi-stream HMMs that are widely used in audio-visual speech recognition, it is important to automatically and properly adjust stream weights. This paper proposes a stream-weight optimization technique based on a likelihood-ratio maximization criterion. In our audio-visual speech recognition system, video signals are captured and converted into visual features using HMM-based techniques. Extracted acoustic and visual features are concatenated into an audio-visual vector. A multi-stream HMM is obtained from audio and visual HMMs. Experiments are conducted using Japanese connected digit speech recorded in real-world environments. Applying the MLLR (maximum likelihood linear regression) adaptation and our optimization method, we achieve a 29% absolute accuracy improvement and a 76% relative error rate reduction compared with the audio-only scheme.

## 1. INTRODUCTION

Automatic speech recognition (ASR) systems are expected to play important roles in user-friendly human-machine interfaces in the near future, such as under ubiquitous computing environments. Although high recognition accuracy can be obtained for clean speech, accuracy dramatically decreases in noisy conditions. Increasing robustness is one of the most important challenges currently for ASR. Multimodal ASR which jointly uses acoustic and visual features has recently become very attractive for this purpose [2, 3, 4]. In most of the multi-modal ASR methods, multi-stream HMMs are used in order to effectively combine acoustic and visual information. An audio-visual multi-stream HMM has audio and visual feature streams, and an audio-visual likelihood is computed as a summation of audio and visual likelihoods weighted by stream weights. Although the stream weights need to be properly estimated according to noise conditions, theoretically they cannot be determined based on the maximum likelihood criterion. Therefore, it is very important to build an efficient stream-weight optimization technique to achieve high recognition accuracy.

This paper proposes an automatic stream-weight optimization method for audio-visual speech recognition based on a likelihood-ratio maximization criterion, in which differences of log likelihood values obtained from the first and other hypotheses are maximized. We evaluate the robustness of our proposed method by conducting experiments using real-world audio-visual data.

In Section 2, we explain the proposed stream-weight optimization method. Our ASR system is described in Section 3, and the experimental setup and results are described in Section 4. Finally, Section 5 concludes this paper.

## 2. STREAM WEIGHT OPTIMIZATION

### 2.1. Multi-stream HMMs

For our method, we use multi-stream HMMs consisting of audio and visual streams. As described in the previous section, multi-stream HMMs have the advantage that they can effectively combine audio and visual information. The log likelihood $b_w(\mathbf{O}_t)$ of an audio-visual feature $\mathbf{O}_t$ for a word $w$ is represented by the following expression (1):

$$b_w(\mathbf{O}_t) = \lambda_{Aw} b_{Aw}(\mathbf{O}_{At}) + \lambda_{Vw} b_{Vw}(\mathbf{O}_{Vt}) \qquad (1)$$

where $t$ is time, and $b_{Aw}(\mathbf{O}_{At})$ and $b_{Vw}(\mathbf{O}_{Vt})$ are likelihoods for an audio feature $\mathbf{O}_{At}$ and a visual feature $\mathbf{O}_{Vt}$, respectively. $\lambda_{Aw}$ and $\lambda_{Vw}$ are audio and visual stream weight factors, respectively, that are constrained by the following restriction (2):

$$\lambda_{Aw} + \lambda_{Vw} = 1 \quad , \quad 0 \leq \lambda_{Aw}, \lambda_{Vw} \leq 1 \qquad (2)$$

### 2.2. Optimization method

As described above, stream weights cannot be determined on the maximum likelihood criterion, in contrast with other model parameters such as mean or variance values of Gaussian components. In the recognition process, these weights need to be estimated properly according to the noise condition in order to achieve high recognition accuracy. Thus, we propose an automatic stream-weight optimization method based on the likelihood-ratio maximization criterion.

Let's denote an output word from a decoder by $w_t$ and a correct word by $\bar{w}_t$ at time $t$. Then the following equation (3) is obtained:

$$b_{w_t}(\mathbf{O}_t) \geq b_w(\mathbf{O}_t) \qquad (3)$$

where $w$ is any word in the dictionary for recognition $W$ ($|W| = N$). A recognition error, that is $\bar{w}_t \neq w_t$, is caused by mismatch between training and testing conditions, making the likelihood of an incorrect word $w_t$ larger than that of the correct word. However, if a majority of the spoken utterances for adaptation are correctly recognized, by adjusting stream weight factors to maximize the difference between the likelihood values obtained from the first and other hypotheses, recognition errors can be expected to decrease. In our method, the set of audio stream weights $\Lambda = \{\lambda_{Aw}\}$ are adjusted to maximize the following equation (4):

$$L(\Lambda) = \sum_{t=1}^{T} \sum_{w \in W} \Big\{ b_{w_t}(\mathbf{O}_t) - b_w(\mathbf{O}_t) \Big\}^2 \tag{4}$$

where $T$ is the total length of the adaptation data. For any $\lambda_{Aw_r} \in \Lambda$, the partial derivative of $L(\Lambda)$ should be zero:

$$\frac{\partial L(\Lambda)}{\partial \lambda_{Aw_r}} = 0 \tag{5}$$

Then the variation of $\lambda_{Aw_r}$, denoted by $\Delta\lambda_{Aw_r}$, can be calculated as follows:

$$\Delta\lambda_{Aw_r} = \frac{A}{B} \tag{6}$$

$$A = \sum_{\substack{t=1 \\ w_t = w_r}}^{T} \Big\{ N b_{w_r}(\mathbf{O}_t) - \sum_{w \in W} b_w(\mathbf{O}_t) \Big\} + \sum_{\substack{t=1 \\ w_t \neq w_r}}^{T} \Big\{ b_{w_r}(\mathbf{O}_t) - b_{w_t}(\mathbf{O}_t) \Big\} \tag{7}$$

$$B = \sum_{\substack{t=1 \\ w_t = w_r}}^{T} N d_{w_r}(\mathbf{O}_t) + \sum_{\substack{t=1 \\ w_t \neq w_r}}^{T} d_{w_r}(\mathbf{O}_t) \tag{8}$$

$$d_w(\mathbf{O}_t) = b_{Aw}(\mathbf{O}_{At}) - b_{Vw}(\mathbf{O}_{Vt}) \tag{9}$$

All $\lambda_{w_r}$ values are updated at once after obtaining all the variations. A set of the optimized stream weights $\hat{\Lambda}$ is obtained after iterating the process.

## 3. AUDIO-VISUAL ASR SYSTEM

### 3.1. Feature extraction

Figure 1 shows the structure of our audio-visual ASR system. The speech signal is recorded at a 16kHz sampling rate, and a speech frame with a length of 25ms is extracted at every 10ms. Each frame is converted into an acoustic vector consisting of 12-dimensional mel-frequency cepstral coefficients (MFCCs), normalized log energy, and their first and second order derivatives. The cepstral mean normalization (CMN) technique is applied to the MFCCs, and the static log energy is removed. As a result, a 38-dimensional acoustic feature is obtained.

Video sequences are captured at 15Hz sampling rate with a resolution size of 360×240. At first, a contour extraction filter is applied to an input image. A smooth contour is modeled by the following equation (10) and positive values $A_i$ and $B_i$ are simultaneously estimated for each column.

$$v_i(y) \simeq \left| A_i (y - y_0) e^{-B_i (y - y_0)^2} \right| \tag{10}$$

where $i$ is the column number, $v_i(y)$ is a contour value at $(i, y)$ and $y_0$ is the center-of-gravity point for the $i$-th column. Since an integral value of $v_i(y)$ becomes large when a part of a person's lips is contained in the column, the horizontal central coordinate of a mouth, denoted by $C$, is obtained by the following equation (11):

$$C = \sum_{i=0}^{W-1} i \times \int_{-\infty}^{\infty} v_i(y) dy \simeq \sum_{i=0}^{W-1} i \times \frac{A_i}{B_i} \tag{11}$$

Next, an equalization filter and a HSI (Hue, Saturation and Intensity) conversion are applied to the input image. For each column of the image, an 8-dimensional vector, consisting of sine and cosine values of hue, saturation, intensity and their derivatives, is generated by scanning the column from top to down. The width and height of the mouth are measured using these vectors by applying HMM-based forced-alignment and one-path-DP-matching techniques. Figure 2 illustrates a summary of the algorithm. The following five HMMs having three states and eight Gaussian pdfs in each state are built using the Baum-Welch algorithm: upper-skin (US), upper-lip (UL), mouth (M), lower-lip (LL) and lower-skin (LS) HMMs. The height $h$ of the mouth is obtained from the positions in the $C$-th column corresponding to the beginning and ending of the mouth HMM given by the forced alignment technique. In order to detect the mouth and lips, the following three scores are computed for each column using the HMMs described above; likelihoods for (a) having lips and a mouth (US → UL → M → LL → LS), (b) having lips (US → UL → LL → LS), and (c) having skin only (US → LS). The one-path DP matching is performed from left to right in the image in order to find the path which maximizes the summation of the scores. The width $w$ of the mouth is obtained from a detected mouth area (a) by using a back track technique. Additionally, by applying a B/W filter to the area between the upper and lower lips in the $C$-th column, teeth information $t$ is obtained by counting detected white pixels. Finally, a 9-dimensional visual feature vector consisting of a parameter set $(h, w, t)$ and its first and second order derivatives is obtained.

After normalizing the dynamic range, the visual vectors are interpolated from 15Hz to 100Hz by a 3-degree spline function so that the frame rate synchronizes with that of the audio vectors. The acoustic and visual features are concatenated to build a 47-dimensional audio-visual vector.

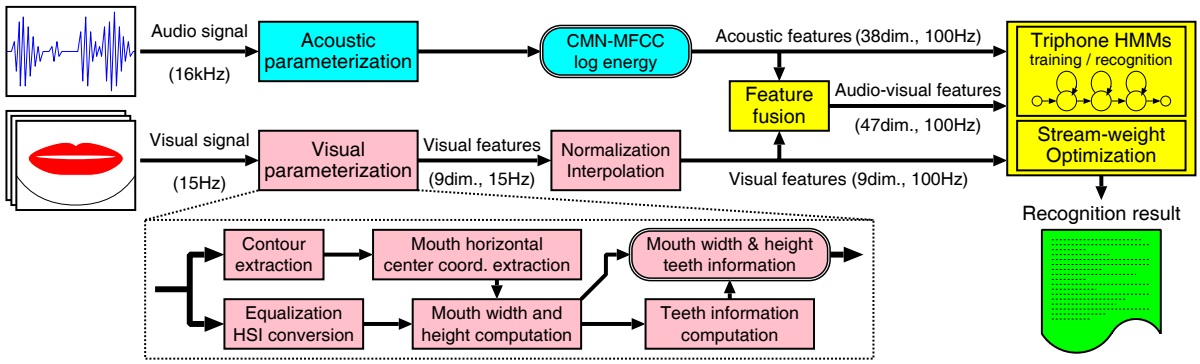


**Fig. 1**. Proposed audio-visual speech recognition system.

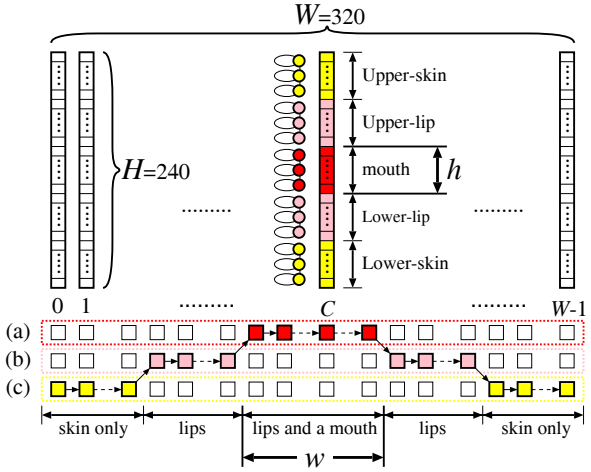


**Fig. 2**. A summary of measuring width and height of a mouth using HMMs.

**Table 1**. Recognition conditions according to audio-visual features, MLLR, and stream-weight optimization.

| | Parameter | MLLR | Optimization |
|---|---|---|---|
| (1) | Audio-only | no | no |
| (2) | Audio-visual | no | no |
| (3) | Audio-visual | no | yes |
| (4) | Audio-visual | yes | no |
| (5) | Audio-visual | yes | yes |

### 3.2. Modeling

A triphone HMM having three states and two mixtures in each state is used in our system. The audio and visual HMMs are built sequentially [6]. First, an audio HMM is trained for the acoustic features, and the phoneme segment information (labels) for the training data is obtained by the forced-alignment technique using the audio HMM. A visual HMM is then built for visual features using the phoneme labels. Finally, the audio and visual HMMs are combined to build an audio-visual multi-stream HMM.

**Table 2**. The highest recognition accuracy in each experimental condition.

| (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|
| 62.0% | 64.2% | 76.1% | 85.2% | 91.0% |

## 4. EXPERIMENTS

### 4.1. Database

Two audio-visual speech databases were collected for training and testing [5]. The first database for training was collected in a clean condition. This database consisted of 2,750 utterances by 11 speakers, each uttering 250 sequences of 2-6 connected Japanese digits. The second database for testing was collected in a driving car on expressways. This consisted of 690 utterances by six speakers, each uttering 115 sequences. There exist several kinds of acoustic and visual noises in this database: engine sounds, wind noises, blinker sounds as acoustic noises, and extreme brightness changing, head shaking on bumpy roads and slow car-frame shadow movements as visual noises.

### 4.2. Adaptation and stream-weight optimization

Table 1 shows the experimental conditions. Experiments were conducted using only acoustic features as a baseline and then using audio-visual features for evaluating the proposed method. When using audio-visual vectors, a common audio and visual stream weight was first applied to all the HMMs. We applied either the MLLR adaptation technique [1] or our proposed stream-weight optimization method, or both of them. The MLLR adaptation was conducted only for the audio stream in an unsupervised batch adaptation manner to adapt mean and variance values. Optimized stream weights were obtained with 50 iterations using the whole testing data and time-aligned labels generated from the recognition results using the common stream weight. In the case of condition (5), the MLLR adaptation was applied first, followed by the stream-weight optimization.

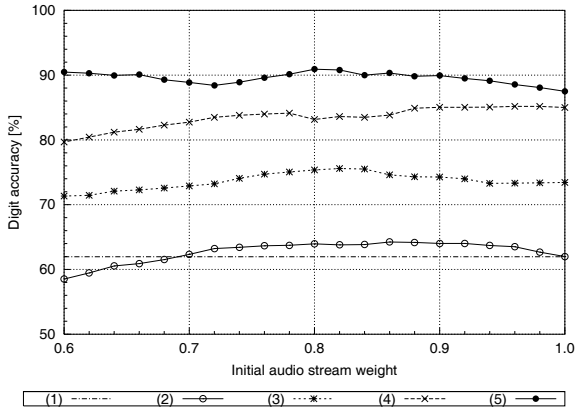


**Fig. 3**. Digit recognition accuracy by each method as a function of the initial audio stream weight.

**Table 3**. Comparison of the number of each class of recognition errors: deletion, substitution and insertion.

| | Deletion | Substitution | Insertion |
|---|---|---|---|
| (1) | 226 | 613 | 211 |
| (3) | 162 | 277 | 149 |

### 4.3. Experimental results

Figure 3 shows the recognition results as a function of the initial audio stream weight used in the iterative process of stream-weight optimization. The horizontal axis indicates the initial audio stream weight, and the vertical axis indicates the digit recognition accuracy. Table 2 shows the highest recognition accuracy by each method. By comparing the results in conditions (1) and (2), it can be seen that approximately 2% absolute improvement was achieved by combining the visual information. By applying the stream-weight optimization (3), a 14% improvement of digit accuracy from the baseline and a 37% relative reduction of digit error rate were obtained. These results indicate the effectiveness of the proposed stream-weight optimization method for multi-stream HMMs. Comparing the results in conditions (4) and (5), a 6% accuracy improvement and roughly 39% error reduction were achieved by the stream-weight optimization. Based on this result showing that approximately the same error reduction can be achieved irrespective of whether the MLLR adaptation is applied or not, it can be concluded that the stream-weight optimization method is effective in a wide range of recognition conditions. Finally, comparing the results in conditions (1) and (5), a 29% accuracy improvement and a 76% error reduction were achieved by combining the visual information and applying both the MLLR adaptation and the stream-weight optimization. Figure 3 also indicates that results with both the MLLR and the stream-weight optimization (5) are not strongly affected by the initial stream weight. This means that stable high recognition accuracies can be expected by applying the MLLR and the weight optimization methods.

A supplementary analysis was conducted to analyze factors of improvements. Table 3 shows the number of recognition errors in the conditions of (1) and (3) categorized into deletion, substitution and insertion errors. The results show that substitution errors are significantly reduced by using visual features together with the weight optimization method, which means that discrimination power is dramatically increased by the proposed method.

## 5. CONCLUSIONS

This paper has proposed an automatic stream-weight optimization method for multi-modal speech recognition using multi-stream HMMs, and has evaluated the robustness of the proposed method against both acoustic and visual noises using real-world data. Our proposed method in combination with an MLLR unsupervised adaptation achieved a 29% absolute improvement of digit accuracy and a 76% relative reduction of digit error rate in comparison with the audio-only baseline method.

Our future works include: (1) comparing the performance of the proposed method with other methods under real environments, (2) testing of proposed techniques for larger data sets or more difficult tasks, and (3) investigation of fusion algorithms and audio-visual synchronization methods.

## 6. ACKNOWLEDGEMENTS

This research has been conducted in cooperation with NTT DoCoMo Multimedia Laboratories. The authors wish to express their thanks for their support.

## 7. REFERENCES

[1] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, pp.171-185, 1995.

[2] G. Potamianos and E. Cosatto and H.P. Gref and D.B. Roe, "Speaker independent audio-visual database for bimodal ASR," *Proc. AVSP'97*, pp.65-68, 1997.

[3] C. Miyajima and K. Tokuda and T. Kitamura, "Audio-visual speech recognition using MCE-based HMMs and model-dependent stream weights," *Proc. ICSLP2000*, vol.2, pp.1023-1026, 2000.

[4] S. Nakamura and H. Ito and K. Shikano, "Stream weight optimization of speech and lip image sequence for audio-visual speech recognition," *Proc. ICSLP2000*, vol.3, pp.20-24, 2000.

[5] S. Tamura, K. Iwano and S. Furui, "A robust multi-modal speech recognition method using optical-flow analysis," *Proc. IDS02*, Closter Irsee, Germany, pp.2-4, 2002.

[6] T. Yoshinaga, S. Tamura, K. Iwano and S. Furui, "Audio-visual speech recognition using lip movement extracted from side-face images," *Proc. AVSP2003*, St Jorioz, France, pp.117-120, 2003.