Audio-visual Interaction in Model Adaptation for Multi-modal Speech Recognition

Satoshi Tamura, Masanao Oonishi and Satoru Hayamizu
Department of Information Science, Gifu University, Japan
E-mail: {tamura@info., oonishi@asr.info., hayamizu@} gifu-u.ac.jp

***Abstract*—This paper investigates audio-visual interaction, i.e. inter-modal influences, in linear-regressive model adaptation for multi-modal speech recognition. In the multi-modal adaptation, inter-modal information may contribute the performance of speech recognition. Thus the influence and advantage of inter-modal elements should be examined. Experiments were conducted to evaluate several transformation matrices including or excluding inter-modal and intra-modal elements, using noisy data in an audio-visual corpus. From the experimental results, the importance of effective use of audio-visual interaction is clarified.**

I. Introduction

In order to enhance the robustness of Automatic Speech Recognition (ASR) in noisy or real environments, multi-modal speech recognition (or bimodal ASR, audio-visual ASR) is often employed [1], [2], [3]. In the typical multi-modal ASR, speech signals as well as lip images, that are not affected by any acoustic noises, are used together. The multi-modal ASR has achieved the better performance than the conventional audio-only ASR in acoustically noisy environments.

On the other hand, model adaptation technique has been widely used by many ASR methods. In the model adaptation, model parameters in acoustic models are modified so as to decrease the mismatch between the model parameters and adaptation features. Maximum Likelihood Linear Regression (MLLR) [4] is one of the major adaptation methods for speech processing; MLLR updates model parameters such as mean vectors in Gaussian distributions by linear transformation.

In multi-modal ASR, MLLR is also used in order to further improve recognition accuracy, and actually achieving high performances [1]. Since the basic MLLR method adapts model parameters assuming a single modality, there may be mutual influences between modalities in multi-modal ASR: e.g. contributions of audio features to visual adaptation and visual effect to audio model parameters. Regarding adaptation for multi-modal ASR, there is a related work in which several adaptation techniques were used in order to convert features [5]. However, there are few researches investigating inter-modal or intra-modal influences and the audio-visual interaction.

In this paper, we investigate the inter-modal effects in linear-regressive adaptation for audio-visual ASR. Several MLLR transformation matrices are analyzed by evaluating the performance of multi-modal ASR.

This paper is organized as follows: Section II introduces multi-modal speech recognition and its corpus used in this paper. The principle of model adaptation is described in Section III. Experimental setup, result and discussion are shown in Section IV. Finally Section V concludes this paper.

TABLE I
Specification of CENSREC-1-AV.

(A) speech data

Sampling freq.	16 kHz
Bit rate	16 bit/sample
File format	RIFF Waveform Audio (.wav)
Noise	Interior car noises (driving on city road and expressway)

(B) image data

	Color image	Infrared image
Frame rate	29.97 Hz (NTSC)	
Pixel data	24bit RGB color	8bit grayscale
Image size	width 81 pixel × height 55 pixel	
File format	Windows Bitmap Image (.bmp)	
Distortion	Driving simulation (gamma transformation)	—

(C) data set

	Training set	Test set (adaptation set)
# spkr.	20 females and 22 males	26 females and 25 males
# utter.	3,234 utterances	1,963 utterances
Acoustic	clean	clean, city-road noise (6 SNRs), expressway noise (6 SNRs)
Visual	clean/color, clean/infrared	clean/color, clean/infrared, gamma/color

II. Multi-modal speech recognition

A. CENSREC-1-AV

A corpus "CENSREC-1-AV" (CENSREC: Corpora and Environments for Noisy Speech RECognition) is utilized in this paper [6]. CENSREC-1-AV includes not only speech data and mouth pictures but also a baseline system and its result. By comparing the baseline result, we can easily evaluate our own multi-modal ASR system.

B. Data and features

The data specifications are summarized in Table I (A) and (B). There are approximately 5,200 utterances in total, which consists of Japanese connected digits. Speech signal was recorded in office environment. Two kinds of movies were captured for each utterance: color (optical) and infrared pictures. Infrared pictures are helpful when illumination (visible spectrum) condition is "noisy," e.g. in a driving car.

A 39-dimensional acoustic vector consisting of 12 MFCCs, an energy coefficient, and their first and second derivatives is extracted from an audio frame, of which frame length is 25ms. A 30-dimensional visual feature is also computed, that includes 10-dimensional "eigenlip" components [3] and their Δ and $\Delta\Delta$ coefficients. The frame shift of acoustic and visual features is 10ms.

C. Model training and recognition

Hidden Markov Model (HMM) is employed as acoustic, visual, and audio-visual models. Data set specification is summarized in Table I (C). The training set is used for training, then the test set is also utilized for testing. The model training method in this paper is the same as that of CENSREC-1-AV. Each digit HMM had 16 states respectively, and an HMM for silence had three. A multi-stream HMM consisting of an audio stream derived from the acoustic HMM and a visual stream derived from the visual HMM is employed. In this HMM, an output log likelihood is computed as:

$$b_{av}(\mathbf{o}_{av}) = \lambda_a b_a(\mathbf{o}_a) + \lambda_v b_v(\mathbf{o}_v) \tag{1}$$

where $\mathbf{o}_a$ and $\mathbf{o}_v$ are acoustic and visual features respectively, and $\mathbf{o}_{av} = ({\mathbf{o}_a}^T\ {\mathbf{o}_v}^T)^T$. Audio and visual log likelihoods are denoted by $b_a(\mathbf{o}_a)$ and $b_v(\mathbf{o}_v)$, respectively. Finally λ_a and λ_v are stream weighting factors. When recognition, the stream weights are optimized manually under the constraint:

$$\lambda_a + \lambda_v = 1 \tag{2}$$

The audio stream weight λ_a are tested at intervals of 0.1.

CENSREC-1-AV provides a baseline result in several noisy conditions: two in-car noises recorded on city roads and expressways. Every noises are respectively added to clean speech data in a test set, at six SNRs (-5 to 20dB). As visual distortion, gamma transformation is applied to color pictures in the test set in order to simulate car-driving condition. Thus three kinds of visual data are available: clean/color, clean/infrared and gamma/color.

III. Model adaptation

A. MLLR

Maximum Likelihood Linear Regression (MLLR) is widely used in ASR, that can improve the performance particularly in noisy or real environments. In this paper, a simple HMM in which each state has only one Gaussian pdf is considered. Let us denote an N-dimensional average vector of a Gaussian distribution by $\boldsymbol{\mu}$. MLLR projects the mean vector into an adapted vector $\hat{\boldsymbol{\mu}}$ by the following linear regression:

$$\hat{\boldsymbol{\mu}} = H\boldsymbol{\mu} + \mathbf{b} \tag{3}$$

where H is an N-dimensional square matrix and $\mathbf{b}$ is an N-dimensional bias vector. The equation (3) can be rewritten as:

$$\hat{\boldsymbol{\mu}} = W\ \boldsymbol{\xi} \tag{4}$$

where $\boldsymbol{\xi} = (1\ \boldsymbol{\mu}^T)^T$ and $W = (\mathbf{b}\ H)$. The matrix W can be determined using adaptation features [4].

$$W^{(a)} = \begin{pmatrix} b^{(a)}_1 & h^{(a)}_{1,1} & \cdots & h^{(a)}_{1,N_a} \\ \vdots & \vdots & & \vdots \\ b^{(a)}_{N_a} & h^{(a)}_{N_a,1} & \cdots & h^{(a)}_{N_a,N_a} \end{pmatrix} \qquad W^{(v)} = \begin{pmatrix} b^{(v)}_1 & h^{(v)}_{1,1} & \cdots & h^{(v)}_{1,N_v} \\ \vdots & \vdots & & \vdots \\ b^{(v)}_{N_v} & h^{(v)}_{N_v,1} & \cdots & h^{(v)}_{N_v,N_v} \end{pmatrix}$$

Fig. 1. *MLLR transformation matrices for audio-only and visual-only ASRs.*

B. Adaptation in unimodal ASR

The basic MLLR method explained above is applied to audio-only ASR:

$$\hat{\boldsymbol{\mu}}^{(a)} = W^{(a)}\ \boldsymbol{\xi}^{(a)} \tag{5}$$

where $\boldsymbol{\xi}^{(a)}$ is an extended average vector, $W^{(a)} = (\mathbf{b}^{(a)} H^{(a)})$. Similarly, MLLR is also applied to visual ASR (lipreading):

$$\hat{\boldsymbol{\mu}}^{(v)} = W^{(v)}\ \boldsymbol{\xi}^{(v)} \tag{6}$$

where $\boldsymbol{\xi}^{(v)}$ and $\hat{\boldsymbol{\mu}}^{(v)}$ are an extended visual mean vector and its adapted vector respectively, and $W^{(v)} = (\mathbf{b}^{(v)} H^{(v)})$. Figure 1 depicts the matrices $W^{(a)}$ and $W^{(v)}$. In this figure, N_a and N_v indicate audio and visual dimensions respectively.

C. Adaptation in multi-modal ASR

For multi-modal ASR, the conventional MLLR has been used. However, it is not investigated and clarified how the adaptation method should deal with multiple modalities, or how inter-modal information affects the performance; e.g. whether audio information is effective to adapt visual model parameters or not, and visual features contribute audio adaptation or not. In order to further examine the adaptation for multi-modal ASR, therefore, this paper evaluates the following five MLLR schemes (transformation matrices W_1, W_2, W_3, W_4 and W_5 illustrated in Figure 2 and explained in Table II):

1) <u>Conventional audio-visual adaptation</u>
   The conventional MLLR is applied; a full transformation matrix $W_1 = (\mathbf{b}^{(av)} H^{(av)})$ is obtained using 69-dimensional audio-visual features. Then $H^{(av)}$ can be expressed as:
   $$H^{(av)} = \begin{pmatrix} H^{(av)}_{aa} & H^{(av)}_{av} \\ H^{(av)}_{va} & H^{(av)}_{vv} \end{pmatrix} \tag{7}$$
   In the following explanation, let us denote $H^{(av)}_{xy}$ by H_{xy}. In this case, audio adaptation is conducted using not only audio but also visual information. Visual adaptation is also accomplished in the same way.
2) <u>Intra-modal adaptation obtained by multi-modal data</u>
   A transformation matrix W_2 is derived from W_1, however, only intra-modal elements (H_{aa} and H_{vv}) remain and inter-modal elements (H_{av} and H_{va}) are discarded.
3) <u>Intra-modal adaptation obtained by unimodal data</u>
   Similar to W_2, a matrix W_3 has only intra-modal transformation. The audio part is equivalent to the audio-only matrix $H^{(a)}$, and the visual part is as same as the visual one $H^{(v)}$. The difference between W_2 and W_3 is that, the intra-modal elements are computed using audio-visual data in the former, while the elements are obtained using audio and visual data respectively in the latter.
4) <u>Multi-modal audio adaptation / visual-only adaptation</u>
   An audio mean vector is adapted using audio-visual information (H_{aa} and H_{av}), whereas a visual mean

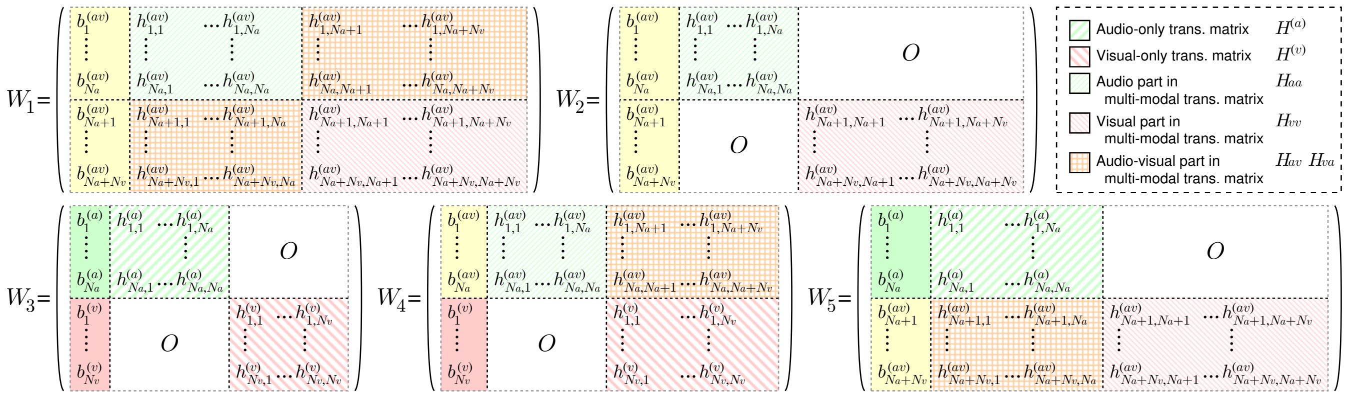

Fig. 2. *MLLR transformation matrices for multi-modal ASR (see also Table II).*

TABLE II
Audio and visual adaptation in MLLR for multi-modal ASR.

	Audio adaptation		Visual adaptation	
W_1	AV	$(H_{aa}\ H_{av})$	AV	$(H_{va}\ H_{vv})$
W_2	AV(a)	(H_{aa})	AV(v)	(H_{vv})
W_3	A	$(H^{(a)})$	V	$(H^{(v)})$
W_4	AV	$(H_{aa}\ H_{av})$	V	$(H^{(v)})$
W_5	A	$(H^{(a)})$	AV	$(H_{va}\ H_{vv})$

AV ··· multi-modal (audio and visual) transformation,
AV(a) ··· only audio part in multi-modal transformation,
AV(v) ··· only visual part in multi-modal transformation,
A ··· audio-only transformation,
V ··· visual-only transformation.

vector is affected only by visual parameters ($H^{(v)}$) in a matrix W_4. By comparing this matrix with W_1 and W_3, inter-modal effect can be further investigated.

5) Audio-only adaptation / multi-modal visual adaptation
A matrix W_5 adapts audio mean parameters using only acoustic data. Visual adaptation is then performed using audio and visual information. This matrix is also designed to analyze the audio-visual interaction.

IV. Experiment

A. Experimental setup

In the following experiments, a simple HMM having only one audio mixture and one visual mixture was employed. Two kinds of models were built using clean audio and color visual data, as well as clean audio and infrared visual data.

For each speaker, the global unsupervised adaptation was applied; one transformation matrix was shared by all states in all HMMs. 10 utterances (equivalent to roughly 30-second utterances) in subject's data in the test set were used for adaptation, all the subject's utterances were then recognized. The mean values were adapted whereas no adaptation was applied for covariance and transition matrices, and mixture weights. Three audio noise conditions were used: expressway 20dB, 10dB and 0dB. All the three visual conditions were employed for testing. Therefore, every adaptation methods were tested in nine audio-visual conditions. Recognition parameters, i.e. an insertion penalty and stream weights, were optimized manually to achieve the best performance for each condition. Any other experimental conditions (features, training, recognition, and noises) are the same as those of CENSREC-1-AV.

TABLE III
Recognition accuracies of audio and visual ASRs.

(A) audio-only ASR (conventional ASR)

	w/o MLLR	MLLR
20dB	92.60%	95.78%
10dB	71.47%	95.05%
0dB	51.61%	91.58%

(B) visual-only ASR (lipreading)

	clean/color	clean/infrared	gamma/color
w/o MLLR	36.02%	37.56%	34.24%
MLLR	35.60%	38.63%	34.49%

B. Experimental result of unimodal ASR

Table III shows recognition accuracies of audio-only and visual-only unimodal ASRs in noisy environments. Recognition results before and after MLLR are listed for comparison. $W^{(a)}$ was used for audio adaptation, and $W^{(v)}$ was used for visual adaptation. According to Table III, it is obvious that the audio-only MLLR is much successful. This phenomenon was caused because the acoustic noises used in the experiments have less magnitudes in the frequency domains that are dominated by speech. In contrast, the advantage of visual adaptation is limited. Since the original accuracy is not sufficient, the adaptation might not work well. Infrared results are slightly better than color ones. This may be because the number of mixtures is insufficient for color pictures: eight mixtures for color whereas one mixture for infrared in CENSREC-1-AV.

C. Experimental result of multi-modal ASR

Table IV represents recognition accuracies of the multi-modal ASR. The first result (0) is obtained without using MLLR, and the other results (1)−(5) are given by MLLR adaptation. The conventional adaptation (1) achieved better performance than the baseline result (0), however, no significant difference is observed when comparing to the result of audio-only MLLR in Table III (A). On the other hand, the following remarkable result is observed; comparing to the result of visual-only MLLR shown in Table III (B), the multi-modal MLLR method using W_1 and $\lambda_V = 1$ achieved better performance of 41−45% recognition accuracy shown in Table V. This means that it is effective to use audio-visual adaptation even in lipreading, and maybe in audio-only ASR in some situations where visual performance is superior to audio one.

Comparing the result (2) with (1), audio-visual interaction plays a certain role in improving the performance. Figure 3

TABLE IV

Recognition accuracies of multi-modal ASR.

	clean/color	clean/infrared	gamma/color
(0) without MLLR adaptation			
20dB	93.13%	93.58%	92.88%
10dB	71.59%	71.59%	71.59%
0dB	51.52%	53.13%	51.74%
(1) MLLR (W_1: conventional adaptation)			
20dB	95.04%	96.08%	95.04%
10dB	92.40%	94.68%	92.29%
0dB	89.53%	92.85%	89.37%
(2) MLLR (W_2: using H_{aa} and H_{vv})			
20dB	92.83%	95.64%	92.65%
10dB	89.44%	93.64%	89.53%
0dB	86.24%	91.39%	86.30%
(3) MLLR (W_3: using $H^{(a)}$ and $H^{(v)}$)			
20dB	96.51%	96.51%	96.34%
10dB	95.41%	95.66%	95.24%
0dB	92.68%	93.24%	92.43%
(4) MLLR (W_4: using H_{aa}, H_{av} and $H^{(v)}$)			
20dB	95.04%	96.11%	95.04%
10dB	92.40%	94.68%	92.29%
0dB	89.53%	92.85%	89.37%
(5) MLLR (W_5: using $H^{(a)}$, H_{va} and H_{vv})			
20dB	96.65%	96.60%	96.51%
10dB	95.56%	95.64%	95.35%
0dB	93.00%	93.22%	92.60%

TABLE V

Visual recognition accuracies ($\lambda_A = 0, \lambda_V = 1$) in the condition (1) in Table IV.

clean/color	clean/infrared	gamma/color
42.44	45.31	41.03

illustrates an absolute value of each element in the matrix W_1. In this figure, white means a small value, and black indicates a large value. Inter-modal effects are easily observed, and such the interaction might engage recognition performance.

The intra-modal adaptation method (3), that is one of the normal model adaptation schemes, is significantly superior to the methods (1) and (2). However, it is also shown in the first paragraph that visual performance can be improved by using audio-visual information, i.e. inter-modal information. Therefore, it is predicted that "in the modality that has relative low accuracy, using the another modality that achieves relatively high performance is crucial and effective."

In order to investigate the prediction, the MLLR schemes (4) and (5) are further examined and compared. The method (5) has slightly better results compared to the method (3): for example, approximately 5% relative error reduction using gamma visual features in SNR=20dB audio condition. Visual performance is not significant as described, however, it is turned out that visual information plays a role in improving audio-visual recognition performance to an extent. This is because the best recongition results are observed when using $\lambda_V = 0.2$ or 0.1. On the other hand, the result of the method (4) is not superior to that of the method (3) and is almost the same as that of the method (1). Note that the audio stream

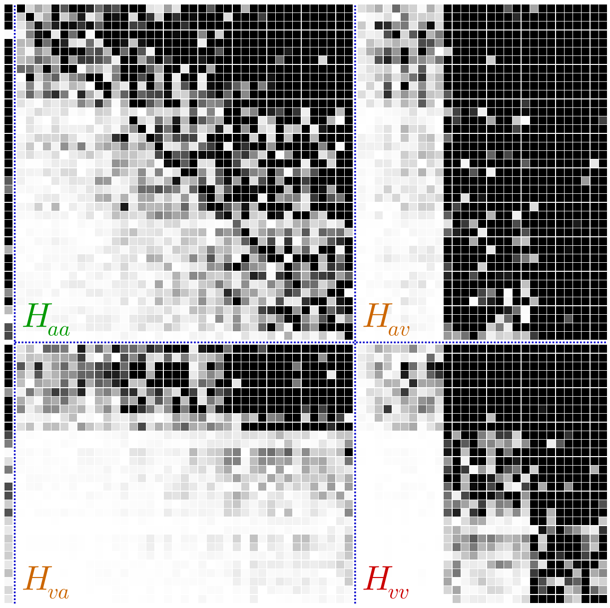

Fig. 3. *A sample of transformation matrix W_1.*

weights in the methods (1) and (4) were almost 1.0, hence, visual adaptation cannot be evaluated by using these results.

V. Conclusion

We investigate audio-visual interaction, or inter-modal influences, in MLLR adaptation for multi-modal speech recognition. Experiments were conducted using several MLLR transformation matrices, the following conclusion is then turned out.

> It is effective and essential for a modality to use the other modalities that have better performance than the modality. It is thus crucial to adopt effective inter-modal information according to conditions of every modalities. And even for a unimodal ASR, there is a great possibility to improve the performance by using the other modalities in adaptation.

Our future work includes automatic stream-weight optimzation for adaptation and recognition, further evaluation of the inter-modal effect using the other corpora, the same investigation to the other adaptation techniques, and development of a high-performance multi-modal ASR system using the results obtained in this paper.

References

[1] S.Tamura et al., "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization," Proc. ICASSP2005, vol.1, pp.469-472 (2005).

[2] G.Potamianos et al., "Hierarchical discriminant features for audio-visual LVCSR," Proc. ICASSP2001, vol.1, pp.165-168 (2001).

[3] C.Miyajima et al., "Audiovisual speech recognition using MCE-based HMMs and model-dependent stream weights," Proc. ICSLP2000, vol.2, pp.1023-1026 (2000).

[4] C.J. Leggetter et al., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, pp.171-185 (1995).

[5] J.Huang et al., "Rapid feature space speaker adaptation for multi-stream HMM-based audio-visual speech recognition," Proc. ICMI2005, pp.338-341 (2005).

[6] S.Tamura et al., "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," Proc. AVSP2010, pp.85-88 (2010).