論文 / 著書情報
Article / Book Information

Title	Improvement of audio-visual speech recognition in cars
Author	Satoshi Tamura, Koji iwano, Sadaoki Furui
Journal/Book name	18th international congress on acoustics (ICA2004), Vol. , No. 4, pp. 2595-2598
発行日 / Issue date	2004, 4

IMPROVEMENT OF AUDIO-VISUAL SPEECH RECOGNITION IN CARS

Satoshi Tamura, Koji Iwano and Sadaoki Furui

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

Abstract

For multi-stream HMMs which are used to effectively combine acoustic and visual information, it is important to optimize stream weights automatically and properly in order to improve the performance. This paper proposes a new stream-weight optimization method based on a likelihood-ratio maximization criterion, in which the difference of log likelihood values between the first and other hypotheses is maximized. Experiments are conducted using Japanese connected digit speech recorded in real-world environments. Applying our unsupervised stream-weight optimization technique, we achieved a 14% absolute accuracy improvement compared with the audio-only scheme, and found that the proposed method is more practical than the MCE-GPD method.

1. Introduction

Automatic Speech Recognition (ASR) systems are expected to play important roles in achieving user-friendly human-machine interfaces in the coming advanced multimedia society supported by ubiquitous computing environments [1]. In particular, ASR has recently attracted a great deal of attention as effective interface for in-car applications, such as car navigation and hands-free communication using cell phones. Although high recognition accuracy can be obtained for clean speech, the accuracy dramatically decreases in noisy conditions. Increasing robustness is one of the most important issues in mobile and vehicular environments. Multi-modal speech recognition, 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 systems, multi-stream HMMs are used in order to effectively combine acoustic and visual information. The audio-visual multi-stream HMMs include weighting factors called stream weights. Since the recognition performance depends on the stream weights and they cannot be automatically optimized by the Maximum Likelihood (ML) criterion, it is crucial to develop an efficient weight optimization technique to achieve high recognition accuracy. One of the most popular methods for this purpose is Minimum Classification Error method using Generalized Probabilistic Descent technique (MCE-GPD). Although the MCE-GPD method is widely used not only in multi-modal ASR but also in other research areas, it has several difficulties for practical applications.

This paper proposes a new unsupervised stream-weight optimization method for audio-visual speech recognition based on the likelihood-ratio maximization criterion, in which the difference of log likelihood values between the first and other hypotheses is maximized. Robustness of the proposed method is evaluated by recognition experiments using in-car audio-visual data.

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

2. Stream weight optimization

2.1. Multi-stream HMMs

In our multi-modal ASR method, multi-stream HMMs are used for recognition. Let us denote a word sequence from a decoder by $w_1, w_2, \cdots, w_M$, an ending time of a segment for each word w_i by T_i, and an audio-visual feature sequence in $T_{i-1} \leq t \leq T_i$ by $\mathbf{O}^i$. The average log likelihood $\bar{b}_w(\mathbf{O}^i)$ for a word w is represented by the following expression:

$$\bar{b}_w(\mathbf{O}^i) = \lambda_{Aw}\bar{b}_{Aw}(\mathbf{O}_A{}^i) + \lambda_{Vw}\bar{b}_{Vw}(\mathbf{O}_V{}^i) \quad (1)$$

where $\bar{b}_{Aw}(\mathbf{O}_A{}^i)$ and $\bar{b}_{Vw}(\mathbf{O}_V{}^i)$ are mean log likelihoods for an audio feature sequence $\mathbf{O}_A{}^i$ and a visual feature sequence $\mathbf{O}_V{}^i$, respectively. λ_{Aw} and λ_{Vw} are audio and visual stream weights respectively, that are constrained by the following restriction:

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

2.2. MCE-GPD algorithm

In recognition processes, stream weights in the multi-stream HMMs need to be estimated properly according to noise conditions to achieve high recognition accuracy. However, the stream weights cannot be determined by the ML criterion as described before.

The MCE-GPD method is one of the typical algorithm that can be applied to this optimization [3, 4]. Let us denote a set of audio stream weights by $\Lambda = \{\lambda_{Aw}\}$. For each word w in a dictionary W, a misclassification measure $d_w(\mathbf{O}^i; \Lambda)$ and a loss function

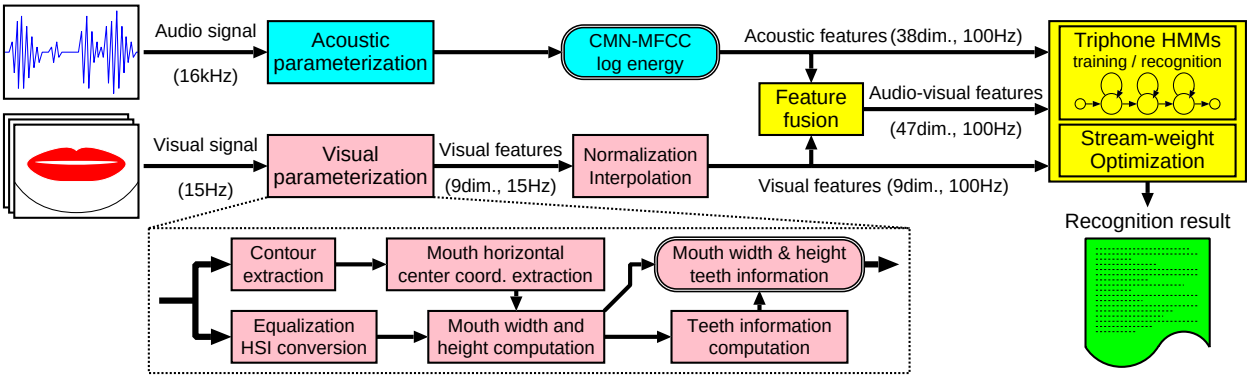

Figure 1: Proposed audio-visual speech recognition system.

$l_w(\mathbf{O}^i;\Lambda)$ are defined as:

$$d_w(\mathbf{O}^i;\Lambda) = -\bar{b}_w(\mathbf{O}^i) + \frac{1}{\eta}\log\left(\sum_{w\in W} e^{\eta\bar{b}_w(\mathbf{O}^i)}\right) \quad (3)$$

$$l_w(\mathbf{O}^i;\Lambda) = \frac{1}{1+e^{-\alpha d_w(\mathbf{O}^i;\Lambda)}} \quad (4)$$

where $\alpha > 0$ and $\eta > 0$ are control parameters. A total loss function is obtained by the following equation:

$$L_M(\Lambda) = \sum_{i=1}^{M} l_{w_i}(\mathbf{O}^i;\Lambda) \rightarrow \min \quad (5)$$

Equation (5) can be minimized in terms of the stream weights by an iterative process using the GPD algorithm:

$$\Lambda_{k+1} = \Lambda_k - \epsilon_k E \nabla L_M(\Lambda_k) \quad (6)$$

where k is an iteration index, ϵ_k is a positive value decreasing as k increases, and E is a unit matrix.

2.3. Proposed optimization method

We propose a new stream-weight optimization method based on the likelihood-ratio maximization criterion. For every word w_i, the following equation can be obtained:

$$b_{w_i}(\mathbf{O}^i) \geq b_w(\mathbf{O}^i) \quad (7)$$

A recognition error is caused by a mismatch between training and testing conditions, making the likelihood of an incorrect word w_i larger than that of the correct word. Therefore, recognition errors are expected to be decreased by adjusting the stream weights so that the difference between likelihood values of the first and other hypotheses is maximized. In our method, the set of audio stream weights Λ is adjusted to maximize the following equation:

$$L_P(\Lambda) = \sum_{i=1}^{M}\sum_{w\in W}\left\{\bar{b}_{w_i}(\mathbf{O}^i) - \bar{b}_w(\mathbf{O}^i)\right\}^2 \rightarrow \max \quad (8)$$

For each word $v \in W$, the following equation is obtained:

$$\frac{\partial L_P(\Lambda)}{\partial \lambda_{Av}} = 0 \quad (9)$$

The amount of variation of λ_{Av}, denoted by $\Delta\lambda_{Av}$, can be calculated as follows:

Table 1: The acoustic and visual feature sets

Acoustic	Frame length	25ms
	Frame period	10ms
	Feature set (38 dim.)	12-dim. CMN-MFCCs, Δ and $\Delta\Delta$ CMN-MFCCs, Δ and $\Delta\Delta$ log power
Visual	Frame period	10ms
	Feature set (9 dim.)	Width of the mouth w, Height of the mouth h, Teeth information t, Δ and $\Delta\Delta$ (w, h, t)

$$\begin{aligned}
\Delta\lambda_{Av} &= \frac{P}{Q} \quad (10)\\
P &= \sum_{i=1}^{M}\Big[\delta_{w_i=v}\cdot\Big\{N\bar{b}_v(\mathbf{O}^i) - \sum_{w\in W}\bar{b}_w(\mathbf{O}^i)\Big\}\\
&\quad + \delta_{w_i\neq v}\cdot\Big\{\bar{b}_v(\mathbf{O}^i) - \bar{b}_{w_i}(\mathbf{O}^i)\Big\}\Big]\\
Q &= \sum_{i=1}^{M}\Big[\delta_{w_i=v}\cdot N\bar{f}_v(\mathbf{O}^i) + \delta_{w_i\neq v}\cdot\bar{f}_v(\mathbf{O}^i)\Big]\\
\bar{f}_w(\mathbf{O}^i) &= \bar{b}_{Aw}(\mathbf{O}_A{}^i) - \bar{b}_{Vw}(\mathbf{O}_V{}^i)
\end{aligned}$$

where N is the total number of words in W. δ_x is 1 if x is true, and 0 otherwise. All $\lambda_{Aw} \in \Lambda$ values are updated at once after obtaining all the variations. A set of optimized stream weights is obtained by iterating this process.

3. Audio-visual ASR system

3.1. Feature extraction

Figure 1 illustrates the structure of our audio-visual ASR system, and Table 1 shows details of the acoustic and visual feature set. Speech signals are recorded at a 16kHz sampling rate, and subsequently a 38-dimensional acoustic vector is obtained for each speech frame.

Video sequences are captured with 360(W) $\times$ 240(H) pixels at 15 frames/sec, recording around speaker's lips. A contour extraction filter is applied to an input image. A smooth contour is modeled by the following equation, in which positive values of A_x and B_x are simultaneously estimated for each column.

$$v(x,y) \simeq \left|A_x(y-y_0)e^{-B_x(y-y_0)^2}\right| \quad (11)$$

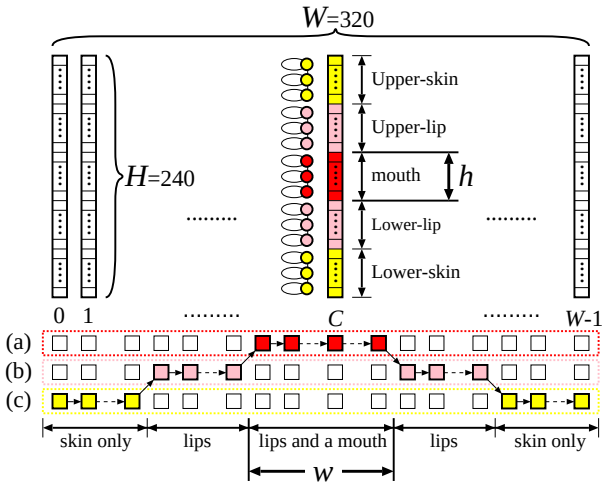

Figure 2: A summary of measuring width and height of a mouth using HMMs.

Here, $v(x, y)$ is a contour value at a point (x, y) and y_0 is the center-of-gravity point for the x-th column. Since an integral value of $v(x, y)$ becomes large when the column crosses the lips area, the horizontal central coordinate of a mouth, denoted by C, is obtained by the following equation:

$$C = \sum_{x=0}^{W-1} \frac{x\, l(x)}{l(x)} \quad \text{where} \quad l(x) = \int_{-\infty}^{\infty} v(x, y)dy \simeq \frac{A_x}{B_x} \quad (12)$$

An equalization filter and an HSI (Hue, Saturation and Intensity) conversion are applied to the input image. For each column of the image, an 8-dimensional vector, consisted 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 the HMM-based forced-alignment and the one-path-DP-matching techniques. Figure 2 illustrates a summary of this 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) · lower-skin (LS)

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. The following three likelihood scores are computed for each column using the above HMMs:

(a) lips and a mouth (US → UL → M → LL → LS)
(b) lips (US → UL → LL → LS)
(c) skin only (US → LS)

The one-path DP matching is performed from left to right in the image in order to find the path that 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 is obtained after the first and second derivatives are computed.

After normalizing the dynamic range, the visual vectors are interpolated 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.

Figure 3: An example of an image in our in-car database (sunlight and car-frame shadow are observed).

3.2. Modeling

Triphone HMMs each having three states and two mixtures in every state is used in our recognition system. The audio and visual HMMs are built sequentially as follows [5]. First, audio HMMs are trained and phoneme labels for the training data are obtained by the forced-alignment technique. Visual HMMs are then built using the phoneme labels. Finally, the audio and visual HMMs are combined into audio-visual multi-stream HMMs.

4. Experiments

4.1. Databases

Two audio-visual speech databases were collected separately for training and testing [6]. The first database, collected in a clean condition, was used for training, in which each one of 11 male speakers uttered 250 sequences of 2-6 connected Japanese digits. The second database, consisting of utterances by six speakers, each uttering 115 sequences, was collected in a driving car on expressways. There exist several kinds of acoustic and visual noises in the latter database: engine sounds, wind noises, blinker sounds as acoustic noises, and extreme brightness changing, head shaking, and slow car-frame shadow movements as visual noises. An example of visual data in the latter database is shown in Figure 3.

4.2. Experimental results

Figure 4 shows recognition results as a function of the initial audio stream weight used in the iterative process of stream-weight optimization for five different methods. 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 obtained by each method. (1) Baseline represents the result using only acoustic features, whereas (2) Audio-visual is the result using audio-visual vectors without optimizing the stream weights. When using

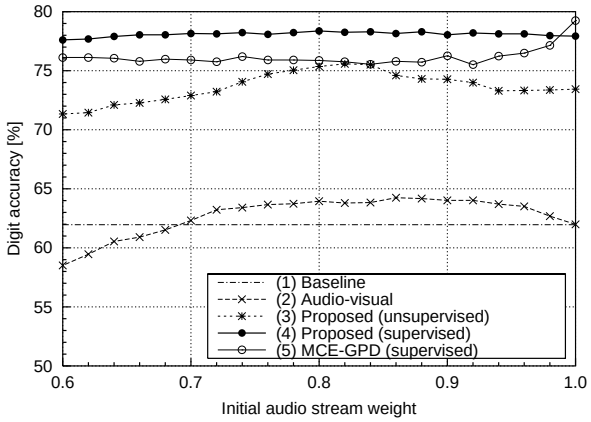

Figure 4: Digit recognition accuracy by each method as a function of the initial audio stream weight.

Table 2: The highest accuracy obtained by each method.

(1)	(2)	(3)	(4)	(5)
62.0%	64.2%	75.6%	78.4%	79.2%

audio-visual vectors, a common audio and visual stream weights were first applied to all the phone HMMs. We applied 50 iterations to both the MCE-GPD and our proposed stream-weight optimization methods. In the (3) Proposed(unsupervised) case, stream weights were adjusted using the whole testing data and time-aligned labels generated from the recognition results. In the (4) Proposed(supervised) and (5) MCE-GPD(supervised) cases, stream weights were optimized on a supervised manner using time-aligned labels obtained from transcriptions. When applying the MCE-GPD, control parameters were varied over $\alpha = (0.8, 1.0, \cdots, 2.0)$, $\eta = (0.8, 1.0, \cdots, 1.4)$, and $\epsilon_k = 1/k$, before determining the optimum parameters that achieved the highest accuracy for each initial stream weight condition.

4.3. Discussions

By comparing the results in conditions (1) and (2), it can be seen that approximately 2% absolute improvement was achieved by simply combining the visual information. By applying our stream-weight optimization (3), a 14% improvement of digit accuracy from the baseline and a 36% relative reduction of digit error rate were obtained. These results indicate effectiveness of the proposed stream-weight optimization method for multi-stream HMMs. Comparing the results in conditions (3) and (4), the accuracy by the unsupervised adaptation was degraded only 3% from the supervised optimization. It means that the accuracy can be effectively improved by the proposed method even in the unsupervised optimization, if a majority of the spoken utterances for adaptation were correctly recognized. Finally we discuss the comparison of conditions (4) and (5). Although the MCE-GPD-based approach achieved a slightly better result than our proposed method at the highest accuracy condition ($\lambda_A = 1.0$, $\alpha = 1.8$, and $\eta = 1.2$), our method exceeded the MCE-GPD in other initial stream weight conditions. From the results shown in Figure 4, the accuracy for the condition (4) was insensitive to initial stream weights. For the MCE-GPD method, it is often difficult to determine proper control parameters that achieve the best recognition accuracy. Hence it can be concluded that our stream-weight optimization method is practical, and the proposed method is expected to be applicable to various real-world tasks.

5. Conclusions

This paper has proposed an automatic stream-weight optimization method based on a likelihood-ratio maximization criterion for multi-modal ASR using multi-stream HMMs. Robustness of the proposed method has been evaluated against both acoustic and visual noises using in-car data. Using our visual feature set and the proposed unsupervised optimization method, we achieved a 14% improvement of digit accuracy and a 36% relative reduction of digit error rate in comparison with the audio-only scheme. We have also found that our optimization method is more practical than the MCE-GPD method.

Our future works include testing of proposed techniques for larger data sets and more difficult tasks, and investigation of fusion algorithms as well as 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] S. Furui, K. Iwano, C. Hori, T. Shinozaki, Y. Saito and S. Tamura, “Ubiquitous speech processing,” Proc. ICASSP2001, vol.1, pp.13–16 (2001–5).

[2] G. Potamianos, E. Cosatto, H.P. Gref and D.B. Roe, “Speaker independent audio-visual database for bimodal ASR,” Proc. AVSP’97, pp.65-68 (1997–9).

[3] C. Miyajima, 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 (2000–10).

[4] S. Nakamura, 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 (2000–10).

[5] T. Yoshinaga, S. Tamura, K. Iwano and S. Furui, “Audio-visual speech recognition using lip movement extracted from side-face images,” Proc. AVSP2003, pp.117-120 (2003–9).

[6] S. Tamura, K. Iwano and S. Furui, “A robust multi-modal speech recognition method using optical-flow analysis,” Proc. IDS02, pp.2-4 (2002–6).