T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title(English)	Toward robust multimodal speech recognition
Authors(English)	Satoshi Tamura, Koji Iwano, Sadaoki Furui
Citation(English)	Symposium on Large Scale Knowledge Resources (LKR2005), Vol. , No. , pp. 163-166
発行日 / Pub. date	2005, 3

Powered by T2R2 (Science Tokyo Research Repository)

Proc. Proc. Symposium on Large-Scale Knowledge Resources (LKR 2005), Tokyo, pp. 163-166 (2005)

TOWARD ROBUST MULTIMODAL SPEECH RECOGNITION

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
E-mail: {tamura,iwano,furui}@furui.cs.titech.ac.jp

ABSTRACT

In this paper, a robust multimodal speech recognition system is proposed in order to improve the performance of automatic speech recognition (ASR). A visual feature extraction technique for real-world data is developed and implemented. Multi-stream hidden Markov models (HMMs) including weighting factors are used to combine audio and visual information, applying the stream-weight optimization scheme based on an output likelihood maximization criterion. The proposed system is evaluated using Japanese connected digit speech recorded in real environments. Using about 10 seconds speech data for stream-weight optimization, a 30% relative error reduction is achieved compared to the result before optimization. By additionally applying the noise adaptation, roughly 60% relative error reduction is obtained over the audio-only scheme using a small amount of optimization data.

1. INTRODUCTION

To function with prime efficiency, the 21st century, which has been called the century of knowledge resources, will necessitate construction and use of large-scale knowledge resources in every domain of research, education and life [1]. ASR is expected to play important roles and to widely contribute to the domain of large-scale knowledge resource processing, such as multimedia database systems and information retrieval systems.

Although high recognition accuracy of ASR can be obtained for clean speech, the accuracy dramatically decreases in noisy conditions or real environments. Increasing robustness is thus one of the most important challenges for current ASR. Multi-modal ASR which jointly uses acoustic and visual features has recently become very attractive for this purpose. Many multimodal ASR schemes have been developed and achieved better performance than when only audio information was used. However, they were evaluated in both acoustically and visually clean conditions. The problem associated with the robustness of multimodal ASR still remains.

This paper proposes a multimodal speech recognition system which is significantly robust to not only acoustic but also visual noises or distortions. In order to further improve the performance of multimodal ASR, an automatic stream-weight optimization method for multi-stream HMMs is also proposed. Real-world audio-visual databases are collected to evaluate the system and optimization algorithm through recognition experiments.

This paper is organized as follows: in Section 2, the proposed multimodal ASR system is introduced. The stream-weight optimization for multi-stream HMMs is mentioned in Section 3. The experimental setup and results are described in Section 4. Finally, Section 5 concludes this paper.

2. AUDIO-VISUAL ASR SYSTEM

2.1. Feature extraction

Figure 1 shows the structure of the proposed audio-visual ASR system. A speech frame with a length of 25ms is extracted every 10ms, and converted into a 38-dimensional audio vector; 12-dimensional mel-frequency cepstral coefficients (MFCCs), normalized log energy, and their first and second order derivatives are computed, then the cepstral mean normalization (CMN) technique is applied to the MFCCs and the static log energy is removed.

Video sequences are captured at a 15Hz sampling rate with a resolution size of 360×240 with 24bit color depth. From each image frame, a 9-dimensional visual vector is extracted by the following processes.

Mouth horizontal center computation After a contour extraction filter is applied to an input image, smooth contours are modeled by equation (1) and positive values A_i and B_i are simultaneously estimated for each column.

$$v_i(y) \simeq \left| A_i(y - y_0)e^{-B_i(y-y_0)^2} \right| \qquad (1)$$

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 part of a person's lips is contained in the column, the horizontal central coordinate of a mouth, denoted by C_t, is obtained by the following equation (2):

$$C_t = \sum_{i=0}^{W-1} i \times \int_{-\infty}^{\infty} v_i(y)dy \simeq \sum_{i=0}^{W-1} i \times \frac{A_i}{B_i} \qquad (2)$$

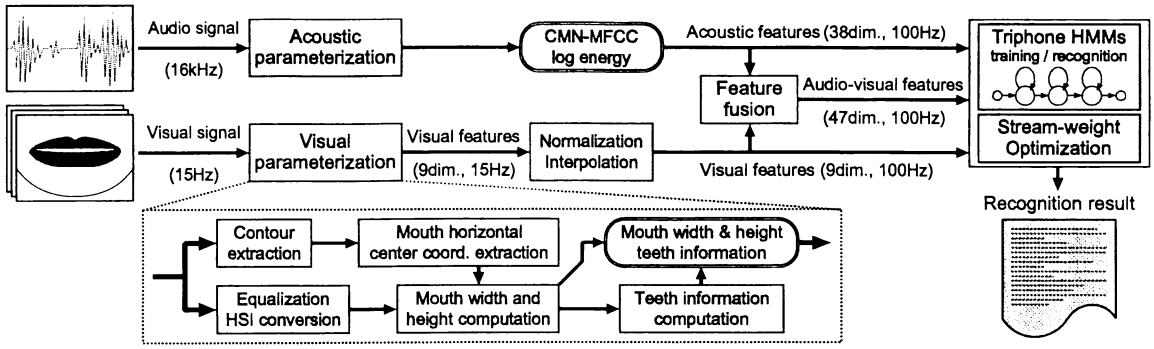

Fig. 1. Principle of proposed audio-visual speech recognition system.

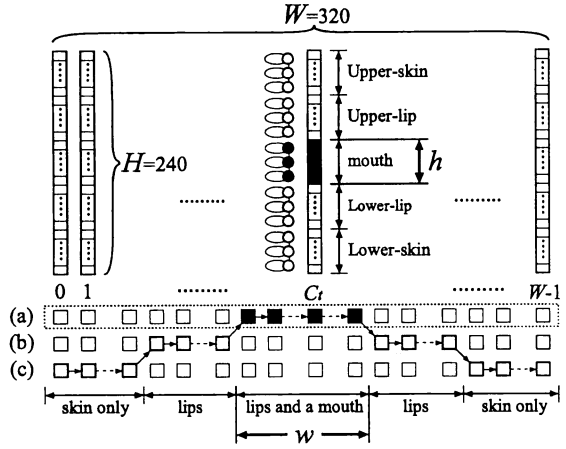

Fig. 2. A summary of measuring width and height of a mouth using HMMs.

Parameter extraction for mouth detection 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, consisting of sine and cosine values of hue, saturation, intensity and their derivatives, is generated by scanning the column from top to bottom.

Mouth width and height computation The width and height of a mouth are measured using these vectors by applying the 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_t-th column corresponding to the beginning and ending of the mouth HMM given by the forced alignment technique. In order to detect horizontal mouth and lip regions, the following three scores are computed for each column using the HMMs described above; likelihoods for

(a) lips and a mouth (US → UL → M → LL → LS),

(b) lips (US → UL → LL → LS), and

(c) skin only (US → LS).

The one-path DP matching is performed from left to right in the image 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 the back-track technique.

Teeth information computation By applying a binary filter to the area between the upper and lower lips in the C_t-th column, teeth information t is obtained by counting detected white pixels. Finally, a parameter set (h, w, t) and its first and second order derivatives are used as a visual feature set.

After synchronizing the frame rates of the audio and visual features to 100Hz, they are concatenated to build a 47-dimensional audio-visual vector and used for recognition.

2.2. Modeling

A triphone HMM having three states and two mixtures in each state is used for speech recognition. The audio and visual HMMs are built sequentially [2]; an audio HMM is trained for the audio 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 labels.

The audio and visual HMMs are combined to build an audio-visual multi-stream HMM. Multi-stream HMMs have the advantage that they can effectively combine audio and visual information. In an audio-visual multi-stream HMM, the log likelihood $b_w(\mathbf{O}_t)$ of an audio-visual feature $\mathbf{O}_t$ for a word $w \in W$ is represented by equation (3):

$$b_w(\mathbf{O}_t) = \lambda_{Aw} b_{Aw}(\mathbf{O}_{At}) + \lambda_{Vw} b_{Vw}(\mathbf{O}_{Vt}) \qquad (3)$$

where t is time, W is a word set of recognition dictionary ($|W| = N$), 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. λ_{Aw} and λ_{Vw} are audio and visual stream weight factors, respectively, that are constrained by the following restriction (4):

$$\lambda_{Aw} + \lambda_{Vw} = 1 \quad , \quad 0 \leq \lambda_{Aw}, \lambda_{Vw} \leq 1 \qquad (4)$$

2.3. Decoding

A Viterbi-based decoder is conducted using the audio-visual features and the multi-stream HMMs, and a word hypothesis is obtained. Stream-weight optimization is performed in an unsupervised manner using the hypothesis, then the weights in the multi-stream HMMs are updated. Finally, a recognition result is obtained by conducting the multimodal speech recognition once again.

3. STREAM-WEIGHT OPTIMIZATION

When recognizing speech data, stream weights in multi-stream HMMs need to be estimated properly according to noise conditions in order to achieve high recognition accuracy. However, the stream weights cannot be determined by the traditional maximum likelihood criterion. In this paper, a new stream-weight optimization method based on an output likelihood normalization criterion is proposed to perform the online optimization.

When there is a mismatch between training and testing conditions, such as in noisy speech recognition, it is often observed that likelihood values of some specific models always become higher or lower than any other model, and this causes recognition errors. For example, if likelihood values of a specific model are always low, the model is hardly selected as recognition results. If dynamic ranges of the models are normalized, all the models have more chance to be selected as a recognition result. In the proposed method, the audio stream weight for a word r can be computed by the following equation (5):

$$\lambda_{Ar} = \frac{\frac{1}{NT}\sum_{t=1}^{T}\sum_{w\in W} b_{Aw}(\mathbf{O}_{At})}{\frac{1}{T}\sum_{t=1}^{T} b_{Ar}(\mathbf{O}_{At})} \qquad (5)$$

where T is the total length of adaptation data. In equation (5), the denominator is the average of log likelihoods for the optimization data obtained from the HMM for the word r, whereas the numerator is the average over all words. Thus equation (5) means that output likelihood values for every word hypothesis are normalized according to the average of the values calculated over the duration of the optimization data including different input words. Each audio stream weight is normalized using the maximum value, before calculating a visual stream weight by equation (4).

4. EXPERIMENTS

4.1. Databases

Two audio-visual speech databases were collected for training and testing [3]. The task of both databases was recognizing Japanese connected digits, each having 2-6 digits, such as "3029 (*san-zero-nī-kyū*)" and "187546 (*ichi-hachi-nana-gō-yon-roku*)". 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 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, blinker sounds as acoustic noises, and extreme brightness changing, head shaking on bumpy roads and slow car-frame shadow movements as visual noises.

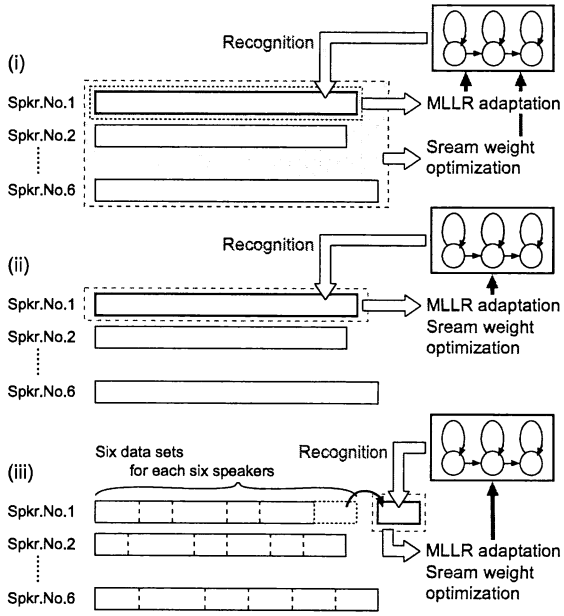

Fig. 3. Recognition conditions for experiments of comparison of both stream-weight optimization methods.

4.2. Experimental setups

Recognition experiments were conducted applying the unsupervised stream-weight optimization. A noise adaptation technique was also conducted in this experiment: maximum likelihood linear regression (MLLR) [4] was applied to the mean and variance values of the audio stream before the stream-weight optimization process.

Speech data spoken by each speaker in the test set were divided into six data sets, yielding 36 data sets in total. The MLLR adaptation and the stream-weight optimization were conducted in the following conditions (also see Figure 3): (i) the stream-weight optimization was conducted using the whole test set, and the MLLR was applied to each speaker; (ii) the optimization and adaptation were conducted for each speaker; and (iii) they were conducted for each one of the 36 data sets.

4.3. Experimental results

Table 1 shows the digit recognition accuracy using audio features or audio-visual features in conditions (i), (ii) and

Table 1. Digit recognition accuracy under various conditions of MLLR adaptation and stream-weight optimization.

MLLR		No		Yes	
S.w. opt.		No	Yes	No	Yes
Audio-only		62.0%		84.9%	
Audio -visual	(i)	64.2%	76.4%	85.1%	90.2%
	(ii)		78.0%		90.4%
	(iii)		77.8%	78.1%	84.5%

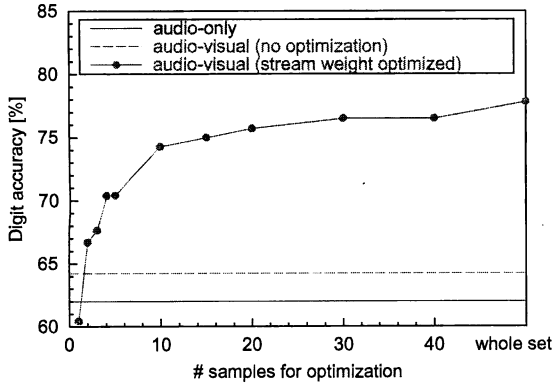

Fig. 4. Digit recognition accuracy as a function of the number of digits used for stream-weight optimization.

(iii) respectively, for every possible combination of the proposed stream-weight optimization and MLLR adaptation. Compared to the audio-only result, approximately 2% absolute improvement of digit recognition accuracy was obtained by using audio-visual features without the MLLR and stream-weight optimization. Applying the stream-weight optimization, approximately 14-16% improvements, equivalent to 38-42% relative reductions of digit error rate, were achieved from the audio-only baseline in all conditions. This means that the performance of the proposed multimodal ASR is greatly improved even in real environments mainly by the stream-weight optimization. When applying both the MLLR adaptation and the stream-weight optimization, it is found that the performance was improved by the MLLR, and further improvements were observed due to the proposed optimization method in all conditions: 74-76% error reductions were obtained in condition (i) and (ii), and a 59% reduction was also achieved in condition (iii) compared to the audio-only result.

The supplemental experiment was conducted to examine performance according to an amount of data for stream-weight optimization. Figure 4 shows the results in condition (iii), as a function of the number of digits used for stream-weight optimization in each data set. For each data set, stream weights were determined using various amounts of digit utterances, and recognition was conducted for whole utterances of the data set. The horizontal axis indicates the number of digits used for optimization, and the vertical axis indicates the digit recognition accuracy. The "whole set" in the horizontal item means that the whole speech data (47-126 digits depending on a set) in each set were used. In Figure 4, it is clearly observed that the more data used, the better the performance of the proposed method becomes. For instance, about 10% absolute accuracy improvement or roughly 30% relative error reduction was achieved compared to the result with no stream-weight optimization, using utterances of only 10 digits, roughly equivalent to 10 seconds of utterances. Therefore, it is concluded that the proposed stream-weight optimization method is effective for online optimization.

These results indicate that the accuracy of the proposed multimodal ASR system was significantly improved using a small amount of data for stream-weight optimization. It can be concluded that the optimization method can estimate the weights properly according to the noise condition of an input data set, and is capable of online stream-weight optimization. Furthermore, by applying the MLLR and optimization, roughly 60% relative digit error reduction was achieved compared to the result of the audio-only method in condition (iii). Hence, it can be concluded that the proposed audio-visual ASR is useful even in real-world environments.

5. CONCLUSIONS

This paper has proposed a multimodal speech recognition system including a stream-weight optimization scheme for multi-stream HMMs. The multimodal ASR system is able to achieve better performance than the audio-only method in real-world data applying the stream-weight optimization, even in the condition where a small amount of optimization data is used. Roughly 60% reduction of recognition error was achieved by combining the proposed audio-visual ASR and MLLR noise adaptation using small data sets.

Our future work includes: (1) investigation of a more effective visual feature set of reduced computational complexity, and (2) testing of the proposed multimodal ASR for more difficult tasks such as large vocabulary continuous speech recognition (LVCSR).

6. REFERENCES

[1] S. Furui, "Overview of the 21st century COE program "Framework for Systematization and application of large-scale knowledge resources,"" *Proc. LKR2004*, Tokyo, Japan, pp.1-8 (2004-3).

[2] 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 (2003-9).

[3] 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).

[4] 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 (1995-4).