論文 / 著書情報
Article / Book Information

Title	A Stream-weight optimization method for multi-stream HMMs based on likelihood value normalization
Author	Satoshi Tamura, Koji Iwano, Sadaoki Furui
Journal/Book name	IEEE ICASSP2005, Vol. SP, No. P5.2, pp. I-469-472
発行日 / Issue date	2005, 3
権利情報 / Copyright	(c)2005 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 MULTI-STREAM HMMS BASED ON LIKELIHOOD VALUE NORMALIZATION

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

In the field of audio-visual speech recognition, multi-stream HMMs are widely used, thus how to automatically and properly determine stream weight factors using a small data set becomes an important research issue. This paper proposes a new stream-weight optimization method based on an output likelihood normalization criterion. In this method, the stream weights are adjusted to equalize the mean values of log likelihood for all HMMs. based on likelihood-ratio maximization which achieved significant improvement by using a large optimization data set. The new method is evaluated using Japanese connected digit speech recorded in real-world environments. Using 10 seconds speech data for stream-weight optimization, a 10% absolute accuracy improvement is achieved compared to the result before optimization. By additionally applying the MLLR (maximum likelihood linear regression) adaptation, a 23% improvement is obtained over 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, the accuracy dramatically decreases in noisy conditions such as driving environments. Thus, increasing robustness is 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 [2, 3, 4]. In most multi-modal ASR methods, multi-stream HMMs are used in order to effectively combine acoustic and visual information. The multi-stream HMM includes audio and visual streams, and weighting parameters called stream weight factors. Although these stream weights need to be properly estimated according to noise conditions, theoretically they cannot be determined by the maximum likelihood (ML) criterion. Therefore, we investigated a stream-weight optimization method based on likelihood-ratio maximization [7]. Significant improvements for recognition accuracy were achieved for real-world data by applying this method; however, a large amount of speech data was needed to robustly determine the stream weights.

In order to optimize these stream weights using a small data set for online audio-visual ASR, this paper proposes a new stream-weight optimization method based on an output likelihood normalization criterion. In this method, stream weights are computed so that output log likelihoods obtained from all multi-stream HMMs are equalized. We compare performance of the proposed optimization scheme with that of the previous optimization method through recognition experiments using real-world audio-visual data.

In Section 2, we explain the details of multi-stream HMMs as well as stream-weight optimization methods. The experimental setups and results are described in Section 3. Finally, Section 4 concludes this paper.

2. STREAM-WEIGHT OPTIMIZATION

2.1. Multi-stream HMMs

In our audio-visual speech recognition scheme, we use multi-stream HMMs consisting of audio and visual streams. 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 the following expression (1):

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

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 (2):

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

2.2. Optimization method based on a likelihood-ratio maximization

Stream weights cannot be determined by the ML criterion, in contrast with other model parameters such as mean or variance values of Gaussian components. When recognizing speech data, these stream weights need to be estimated properly according to noise conditions in order to achieve high recognition accuracy.

We investigated an automatic stream-weight optimization method based on a likelihood-ratio maximization criterion [7]. Recognition errors are caused by a mismatch between training and testing conditions, making the likelihood of an incorrect word larger than that of a correct word. If stream weights could be adjusted to maximize the difference between the likelihood values obtained from the top and other hypotheses, recognition errors could be expected to decrease. In this method, the set of audio stream weights $\Lambda = \{\lambda_{Aw}\}$ are adjusted to maximize the following equation:

$$L(\Lambda) = \sum_{t=1}^{T} \sum_{w \in W} \left\{ b_{w_t}(\mathbf{O}_t) - b_w(\mathbf{O}_t) \right\}^2 \quad (3)$$

where w_t is an output word from a decoder at time t, and T is the total length of adaptation data. From equation (3), the variation of λ_{Av} for a word $v \in W$, denoted by $\Delta\lambda_{Av}$, can be calculated as follows:

$$\Delta\lambda_{Av} = \frac{A}{B} \quad (4)$$

$$A = \sum_{\substack{t=1 \\ w_t = v}}^{T} \left\{ N b_v(\mathbf{O}_t) - \sum_{w \in W} b_w(\mathbf{O}_t) \right\} + \sum_{\substack{t=1 \\ w_t \neq v}}^{T} \left\{ b_v(\mathbf{O}_t) - b_{w_t}(\mathbf{O}_t) \right\}$$

$$B = \sum_{\substack{t=1 \\ w_t = v}}^{T} N d_v(\mathbf{O}_t) + \sum_{\substack{t=1 \\ w_t \neq v}}^{T} d_v(\mathbf{O}_t)$$

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

Finally, the set of optimized stream weights is obtained after iterating this process.

2.3. Optimization method based on an output likelihood normalization

By the optimization method described above, significant improvements of recognition accuracy have been achieved using a large amount of speech data; however, the accuracy sometimes decreases when using a small data set. This degradation is caused by local optimization as a result of the iterative process using a small size of data, or inadequate weights due to lack of optimization data. For real-world applications, it is necessary to develop an online optimization scheme using a small amount of speech data.

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 a more chance to be selected as a recognition result. Hence, we propose a new stream-weight optimization method based on an output likelihood normalization criterion, in which the audio stream weight for a word v can be computed by the following equation:

$$\lambda_{Av} = \frac{\frac{1}{NT} \sum_{t=1}^{T} \sum_{w \in W} \bar{b}_{Aw}(\mathbf{O}_{At})}{\frac{1}{T} \sum_{t=1}^{T} \bar{b}_{Av}(\mathbf{O}_{At})} \quad (5)$$

In the expression (5), the denominator is the average of log likelihoods for optimization data obtained from the HMM for the word v, whereas the numerator is the average over all words. Thus the 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 the equation (2). The proposed method has the advantage that computational complexity is significantly reduced, since no iteration technique is needed.

3. EXPERIMENTS

3.1. Databases

Two audio-visual speech databases were collected for training and testing [5]. 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.

3.2. Audio-visual ASR system

Figure 1 illustrates the structure of our audio-visual ASR system [7]. Acoustic and video signals are recorded using a DV system. The speech signal is converted into a 38-dimensional acoustic vector consisting of 12-dimensional

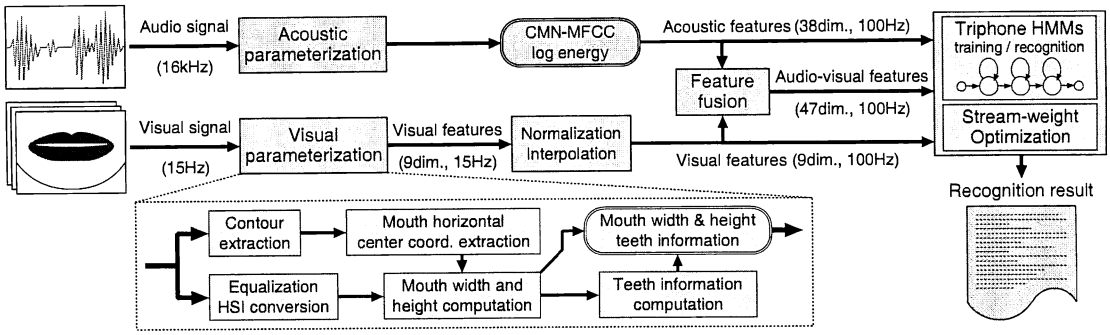

Fig. 1. Principle of our audio-visual speech recognition system.

cepstral-mean-normalized MFCCs and their first and second order derivatives, as well as Δ and $\Delta\Delta$ normalized log energy coefficients. The video signal is converted into a 9-dimensional visual vector consisting of width and height information for a speaker's mouth and teeth, which are measured using HMM-based techniques. After synchronizing the frame rates of the acoustic and visual features, they are concatenated to build a 47-dimensional audio-visual vector and used for recognition.

Audio and visual HMMs are built sequentially [6]; the audio HMM is trained for acoustic features, and then the visual HMM is built for visual features using a phoneme label generated from the acoustic information. Finally, these audio and visual HMMs are combined to build an audio-visual multi-stream HMM.

3.3. Experimental setups

Recognition experiments were conducted in an unsupervised optimization manner, applying the stream-weight optimization based on either the likelihood-ratio maximization (A) or the output likelihood normalization (B). In the case of (A), optimized stream weights were obtained with 50 iterations. The unsupervised maximum likelihood linear regression (MLLR) adaptation [1] 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: condition (i): the stream-weight optimization was conducted using the whole test set, and the MLLR was applied to each speaker, condition (ii): the optimization and adaptation were conducted for each speaker, and condition (iii): they were conducted for each one of the 36 data sets.

3.4. Experimental results

Table 1 shows the digit recognition accuracy using either stream-weight optimization (A) or (B) in condition (i) or

Table 1. Digit recognition accuracy under various conditions of stream-weight optimization (with no MLLR adaptation).

		(A) LRM	(B) OLN
Audio-only		62.0%	
Audio-visual	No opt.	64.2%	
	(i)	75.6%	76.4%
	(iii)	59.4%	77.8%

(A) LRM ... Likelihood-Ratio Maximization
(B) OLN ... Output Likelihood Normalization

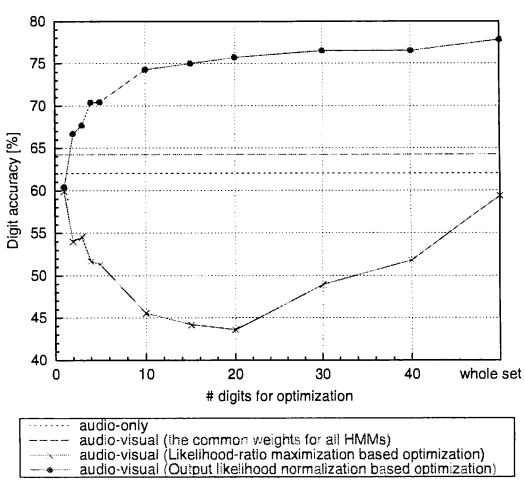

Fig. 2. Digit recognition accuracy as a function of the number of digits used for stream-weight optimization by the proposed and the previous methods.

Table 2. Digit recognition accuracy under various conditions of stream-weight optimization (with MLLR adaptation).

		MLLR only	(A) LRM	(B) OLN
Audio-visual	(i)	85.1%	91.1%	90.2%
	(ii)		88.7%	90.4%
	(iii)	78.1%	76.2%	84.5%

(iii), without the MLLR adaptation. Compared to results obtained without applying the stream-weight optimization, (A) and (B) methods achieved approximately 11% and 12% improvements, respectively, in condition (i). Contrastively, in condition (iii), the proposed method (B) achieved better performance than the results in (i), whereas the accuracy of the previous method (A) became lower than that of the audio-only baseline.

Figure 2 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 according to a set) in each set were used. It is observed that the accuracy by (A) degraded when a small number of digits were used; and on the other hand, the more data are used, the better the performance of (B) becomes.

Finally, Table 2 indicates the recognition results applying both the unsupervised MLLR adaptation and either of stream-weight optimization. The whole speech data in each set were used for the adaptation and optimization. The results show that the performance was improved by the MLLR, and further improvements were observed by applying the proposed method (B) in all conditions. The previous method (A) could not improve performance in condition (iii).

These results indicate that the previous method (A) could not properly determine stream weights when using a small amount of optimization data, and caused degradation of recognition performance. In contrast, the accuracy of the proposed method (B) was significantly improved using a small amount of data. Therefore, it can be concluded that the proposed method optimizes stream weights properly according to the noise condition of an input data set. Experimental results shown in Figure 2 indicate that the proposed method is capable of online stream-weight optimization. For example, about 10% absolute improvement from the result with no stream-weight optimization was achieved using utterances of only 10 digits, roughly equivalent to 10 seconds of utterances. Furthermore, by combining the stream-weight optimization and the MLLR, roughly 23% improvement was achieved compared to the result of the audio-only method, in condition (iii). Hence, it can be concluded that the stream-weight optimization method based on output likelihood normalization is useful, even when the acoustic features are adapted to noise by the MLLR method.

4. CONCLUSIONS

This paper has proposed a new stream-weight optimization method based on an output likelihood normalization criterion for multi-modal speech recognition using multi-stream HMMs. The proposed method can achieve better performance than the previous method for real-world data, especially in the condition where a small amount of optimization data is used. A 23% improvement of recognition accuracy was obtained by combining the stream-weight optimization method and the MLLR using small data sets.

Our future works include: (1) investigation of a more effective visual feature set requiring reduced computational complexity, (2) testing of the proposed techniques for more difficult tasks such as large vocabulary continuous speech recognition (LVCSR), and (3) development of better fusion algorithms and audio-visual synchronization methods.

5. ACKNOWLEDGEMENTS

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

6. 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.

[7] S. Tamura, K. Iwano and S. Furui, "A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs," *Proc. ICASSP2004*, Montreal, Canada, vol.1, pp.857-860, 2004.