STREAM WEIGHT OPTIMIZATION OF SPEECH AND LIP IMAGE SEQUENCE FOR AUDIO-VISUAL SPEECH RECOGNITION

Satoshi Nakamura1 Hidetoshi Ito2 Kiyohiro Shikano2
1ATR Spoken Language Translation Research Laboratories, 2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
2Graduate School of Information Science, Nara Institute of Science and Technology 8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN
E-mail: nakamura@slt.atr.co.jp

ABSTRACT

Bimodal speech recognition systems, with the use of visual information to supplement acoustic information, have been shown to yield better recognition performance than purely acoustic systems, especially when background noise is present. The early integration strategy for HMM-based audio-visual speech recognition is one promising approach, where the output probability is obtained by product of output probabilities of audio and visual streams. This paper addresses a novel method which optimizes stream weights so as to maximize recognition performance. The proposed method estimates the stream weights based on a normalized log likelihood which is derived by ratio of likelihood of a correct word and highest likelihood of incorrect words. The isolated word recognition experiment results show that the audio-visual speech recognition by proposed method attains 56.2 % (10 dB), 55.2 % (0 dB) and 15.2 % (20 dB) better performance compared to that only using audio information. The results also show the proposed method can reduce a number of adaptation words.

1. INTRODUCTION

Speech recognition performance has been drastically improved recently. However, it is also well-known that the performance will be seriously degraded if the system is exposed in noisy environments. Humans pay attention not only to speaker’s speech but also to speaker’s mouth in such adverse environments. The lip reading is the extreme case if it is impossible to get any audio signal. This suggests a fact that speech recognition can be improved by incorporating mouth images. This kind of bi-modal integration is available in almost every situation. Many studies have been presented related to improvements of speech recognition performance in low SNR environments and reduces the number of adaptation words.

As a method for estimating stream weights, the maximum likelihood [4] based method or the GPD [5] based method have been proposed. However, the first method has a serious estimation problem and the latter method has a crucial problem such that many adaptation data is necessary for the weight estimation. In order to cope with these problems, we proposed a new weight estimation method using only a few adaptation data. The isolated word experiments show that the proposed weight estimation improves the recognition performance in low SNR environments and reduces the number of adaptation words.

2. INTEGRATION EFFECTS OF USING LIP IMAGE SEQUENCE

The bimodal speech recognition has a possibility to improve conventional speech recognition performance incorporating visual lip information. The speech recognition early integration scheme can be described as follows,

\[
b(o_t) = b_a(o_t)^{1 - \alpha_A} \cdot b_v(o_t)^{\alpha_V},
\]

(\(\alpha_A + \alpha_V = 1.0\), \(A\) : Audio, \(V\) : Visual)

where \(b(o_t)\), \(b_a(o_t)\), and \(b_v(o_t)\) are output probabilities at time \(t\) for audio-visual, audio, and visual streams, respectively. \(\alpha_A\) and \(\alpha_V\) are stream weights for audio and visual streams, respectively.

Figure 1. Effects of stream weights in bimodal speech recognition
performance is degraded in acoustically noisy environments, whereas visual information is not. However, visual information itself is insufficient to build speech recognition system since its phonetic discriminative performance is so poor. Furthermore it is necessary to consider visual noise interference.

Figure 1 shows the experiment results of speaker dependent 100 words bimodal isolated word recognition based on the early integration. The curves are obtained by changing a stream weight for audio information. Peaks of recognition rates are observed in between audio-only and visual-only conditions in almost acoustic SNR environments. These peaks indicate the effects of information integration of different type of the information.

This paper addresses the problem how to optimize stream weights for audio and visual streams using a few adaptation data from the test environments.

3. PROPOSED METHOD
This paper proposes a new method which tries to solve the problems noted in the previous section. The method estimates the stream weights so as to maximize recognition accuracy. The estimation is carried out by searching better weights iteratively in two regions. However, since the word accuracy doesn’t have good resolution enough for finding optimal stream weights, the increase of the word accuracy usually stops in only a few iterations. The proposed algorithm uses a normalized likelihood instead of the word accuracy. The proposed algorithm may be thought as one kind of approximation of the minimum classification error estimation. However, the proposed method can reduce a number of adaptation data.

3.1. Normalized log-likelihood
The word accuracy is not good measure to estimate stream weights, since the estimation step stops in a few iterations because of lack of resolution. Therefore in this paper, we use the normalized log-likelihood for estimating stream weights. Normalized log-likelihood, \( L_{NR}(O|M_{w_{correct}}) \) is defined as follows,

\[
L_{NR}(O|M_{w_{correct}}) = \frac{L(O|M_{w_{correct}}) - \max_{w_{incorrect}} L(O|M_{w_{incorrect}})}{L(O|M_{w_{correct}})},
\]

(2)

Figure 2. Normalized log-likelihood

3.2. Weights Estimation Procedure
Fig.3 shows the weight search process in the algorithm. The weights are tied for all phonemes and states. The algorithm is summarized as follows,

1. Prepare adaptation words in testing environments and recognition dictionary used in adaptation. This adaptation dictionary includes adaptation words. The larger size of the adaptation dictionary gives the better resolution for estimation.
2. Recognize the adaptation words using 5 kinds of stream weights shown in Fig.3. HMM transition probabilities and output probabilities are estimated using clean speech.
3. Select one of a lower or higher half of the target weight region comparing the normalized log-likelihoods.
4. Set the weight region selected in previous step. Then iterate from step 2. Stop iteration when the weight region becomes smaller than 0.05.

4. EXPERIMENT OF STREAM WEIGHT ESTIMATION
4.1. Audio-visual speech database
Table 1 shows the specification of the database used in this study. Video recording is performed at a sound-proof room and lighting is set from the front. The head is not fixed but the speaker is requested to attach her back to the seat. Moreover, the speaker is also requested to close a mouth before and after utterance. We observed the difference of lighting conditions, size of lips, and inclination of a face every utterance words, since the video recording was conducted over two or more days.

4.2. Preprocessing of visual information
As stated previously, we observed the difference in lighting conditions and inclination of a face in the recorded video
Table 1. Audio-visual database

<table>
<thead>
<tr>
<th>File format</th>
<th>SGI movie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker</td>
<td>One female speaker</td>
</tr>
<tr>
<td>Utterances</td>
<td>ATR 5240 Japanese words</td>
</tr>
<tr>
<td>Audio</td>
<td>sampling: 16bit, 48 kHz</td>
</tr>
<tr>
<td>Visual</td>
<td>sampling rate: 30 [frames/sec] size: 160×120 color quality: 8 bit RGB</td>
</tr>
</tbody>
</table>

Figure 4. Preprocessing (a) Original (b) Histogram change (c) Normalized lip position.

images. Then the images are preprocessed by histogram normalization and the lip position normalization based on pattern matching with a common key frame. Fig 4(a), (b), and (c) are a recorded original image, a histogram normalized image and a lip position normalized image, respectively.

4.3. Experimental condition

Experiment conditions are shown in Table 2. For each modality the basic coefficients and the delta coefficients are collectively merged into one stream.

Since a frame rate of a video image is $\frac{1}{30}$ and then the ratio of the frame rate is is 1:4. We inserted the four same images so as to synchronize the face image frame rate to the audio speech frame rate.

Then, the images are analyzed by 2-dimensional FFT to extract 6x6 log power 2-D spectrum for bimodal speech recognition.

Table 2. Experimental condition.

<table>
<thead>
<tr>
<th>audio</th>
<th>sampling rate: 12 kHz (downsampled) frame length: 32 msec frame shift: 8 msec pre emphasis: $1 - 0.97z^{-1}$ extract:MFCC16 X, $\Delta 16 X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>visual</td>
<td>1: RGB ⊕ 256 grayscale 2: Histogram normalization 3: Lip position normalization 4: 2D-FFT (256 x 256) Parameter: log power spectrum 35 order + $\Delta 35$</td>
</tr>
<tr>
<td>distribution</td>
<td>Gaussian: 3 Mixture 55 phoneme model</td>
</tr>
<tr>
<td>HMM</td>
<td></td>
</tr>
<tr>
<td>training data</td>
<td>4740 word</td>
</tr>
<tr>
<td>adaptation data</td>
<td>15 words (SNR: 30, 20, 10, 0 dB)</td>
</tr>
</tbody>
</table>

Figure 5. Word recognition rates and normalized log-likelihood for stream weights

4.4. Estimation results

Figure 5 shows the number of correct words and sum of the normalized log-likelihood, $L_{NR}$, changing stream weights to the adaptation data set (15 words) in acoustically noisy environments such as audio SNRs are 30dB, 20dB, 10dB, and 0dB.

Generally the peak of the recognition rate differs between adaptation words and testing words. The more adaptation words we use, the more general the stream weights are, and the more computation is needed. In the experiments 15 words chosen at random from testing environments are used for the estimation.

Fig. 5 also shows that the normalized log-likelihood and the word recognition rates has peaks at the same stream weights, and that the normalized log-likelihood gives finer
The resolution determining the optimal stream weights are varied depending on the size of recognition dictionary used for the estimation. We used 1000 words for the estimation.

Fig. 7 shows bimodal speech recognition rates for various SNR conditions. The proposed method outperforms each of audio-only and visual-only speech recognition system, even in SNR 0dB and Clean conditions. The great improvement is observed in very low SNR conditions. These results evidence the proposed method improves speech recognition rates using only 15 adaptation words in acoustically noisy environments.

5. CONCLUSION

In this paper we propose the method technique which effectively estimates audio-visual stream weights of the early integration scheme by few adaptation data in acoustically noisy environments.

The estimation experiments of stream weights are performed and the validity of this technique is shown by the recognition experiments.

Furthermore, we are going to try to investigate noise robust techniques like spectral subtraction as preprocessing for audio information, influences from visual noise, and comparison with the GPD methods. Also the evaluation on continuous speech is to be considered.

6. REFERENCES

STREAM WEIGHT OPTIMIZATION OF SPEECH AND LIP IMAGE SEQUENCE FOR AUDIO-VISUAL SPEECH RECOGNITION

Satoshi Nakamura¹, Hidetoshi Ito² and Kiyohiro Shikano²

¹ ATR Spoken Language Translation Research Laboratories.
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

² Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN
E-mail: nakamura@slt.atr.co.jp

Bimodal speech recognition systems, with the use of visual information to supplement acoustic information, have been shown to yield better recognition performance than purely acoustic systems, especially when background noise is present. The early integration strategy for HMM-based audio-visual speech recognition is one promising approach, where the output probability is obtained by product of output probabilities of audio and visual streams. This paper addresses a novel method which optimizes stream weights so as to maximize recognition performance. The proposed method estimates the stream weights based on a normalized log likelihood which is derived by ratio of likelihood of a correct word and highest likelihood of incorrect words. The isolated word recognition experiment results show that the audio-visual speech recognition by proposed method attains 56.2 % (10 dB), 55.2 % (0 dB) and 15.2 % (20 dB) better performance compared to that only using audio information. The results also show the proposed method can reduce a number of adaptation words.