IMPROVED BIMODAL SPEECH RECOGNITION USING TIED-MIXTURE HMMS AND 5000 WORD AUDIO-VISUAL SYNCHRONOUS DATABASE

Satoshi NAKAMURA, Ron NAGAI, Kiyohiro SHIKANO
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-01, JAPAN
nakamura@is.aist-nara.ac.jp

ABSTRACT

This paper presents methods to improve speech recognition accuracy by incorporating automatic lip reading. The paper improves lip reading accuracy by following approaches; 1)collection of image and speech synchronous data of 5240 words, 2)feature extraction of 2-dimensional power spectra around a mouth and 3)sub-word unit HMMs with tied-mixture distribution(Tied-Mixture HMMs). Experiments through 100 word test show the performance of 85% by lipreading alone. It is also shown that tied-mixture HMMs improve the lip reading accuracy. The speech recognition experiments are carried out over various SNR integrating audio-visual information. The results show the integration always realizes better performance than that using either audio or visual information.

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 to 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 multi-modal integration is available in almost every situation except telephone applications. Many studies have been presented related to improvements of speech recognition by lip images[1, 2, 3, 4, 5, 6, 8, 10]. Recently HMM becomes popular to integrate multi-modal information. This owes to good HMM tools and common databases. However, HMM requires a large amount of training database. It is very difficult to collect speech and lip image synchronous database large enough to estimate lip image HMMs. This paper describes points such as 1)collection of image and speech synchronous data of 5240 word, 2)feature extraction of 2-dimensional power spectra around a mouth, 3)sub-word unit HMMs with tied-mixture distribution(Tied-Mixture HMMs) and 4)their integration to improve speech recognition. The speech recognition performance is evaluated through isolated word recognition experiments over various SNR. Experiment results show that tied-mixture HMM improves lip image recognition accuracy and that the speech and lip image integration improves speech recognition accuracy under various kinds of SNR environments.

2. AUDIO-VISUAL DATABASE

Audio-visual database is corrected by one male speaker. One male speaker utters 5240 ATR Set-A Japanese words in front of a workstation. Lip image is recorded by the camera (Canon VC-C1) adjusting the speaker's lip outline to camera window. The lighting is so arranged that the lip is lighted balanced by a fluorescent lamp. Speech signal is recorded by uni-directional microphone and digitized in 16bit 12kHz. The image and speech data are simultaneously recorded in AVI format. The frame rate is 8msec for speech and 33.3msec for image(30frames/sec). JPEG image data (160x120) is converted to 8bit gray scale image data. Fig.1 shows examples of the recorded lip image data.

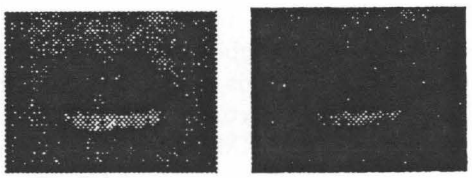

Figure 1. Recorded Images (/a/ left:recorded right:BPF smoothed)

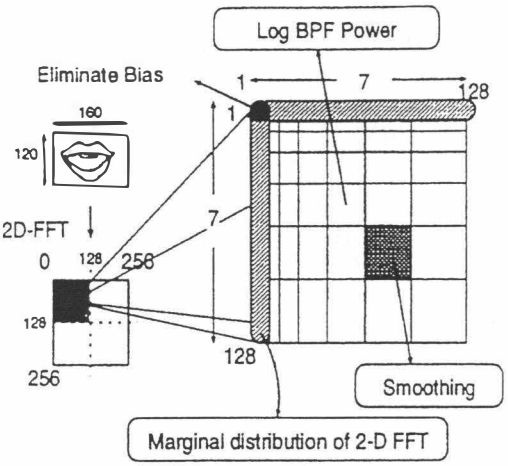

Figure 2. 2-D BPF feature extraction

3. AUTOMATIC LIP READING

Extraction of lip shape characteristics, size of training database and robust modeling of HMMs are quite important issues for automatic lip reading. For feature extraction, there are two approaches such as parametric feature extraction and nonparametric feature extraction. As parametric feature extraction, parameters related to mouth opening are commonly used based on extraction of a mouth shape. To extract the lip shape accurately methods such as deformable templates and active shape models are studied[6, 7, 9, 10]. However, an error of feature extraction causes critical performance degradation by parametric feature extraction. Then in this paper, nonparametric feature extraction is used with sub-word HMMs to avoid the problems. As a nonparametric feature, a gray scale image of lip are analyzed by 2-D FFT shown in fig.2. 2-D log BPF magnitude power are used as nonparametric features.

In 2-D FFT 8bit x 8bit 2-D FFT power spectrum is calculated. The final feature vector is 48 dimensional vector composed of 7x7 2-D log BPF outputs minus the (0,0) term. The right figure in fig.1 shows a BPF smoothed image of the original data. The delta image features over 2 frames are also calculated. Visual HMMs are used to model lip image feature vectors. 55 context independent phoneme HMMs are used. Each model consists of 4 states including initial and final states. The number of words for HMM training is 4740. The number of words for testing are 100 and 500 words. In this paper, Gaussian mixture HMMs(GM) and tied-mixture HMMs(TM) are compared. Tied-mixture HMMs provide robust models for image features under insufficient training data.

Table.1 indicates results of word recognition experiments. The table shows the performance using various number of feature vectors obtained by cutting off higher frequency bands. The 70 dimensional vector shows the best result. It is noticed that linearly spaced BPF is worse than the log spaced bpf in preliminary experiments.

Table 1. Visual speech recognition accuracy vs. number of feature vectors(500 word test; %)

#vectors	Closed		Open	
	GM	Tied-Mix	GM	Tied-Mix
48	65.0	68.0	46.8	49.0
70	69.4	71.4	58.2	60.4
96	60.4	69.6	44.4	47.0

Table 2. Visual speech recognition by marginal distribution with tied-mixture HMMs(%)

	Closed	Open
100	65.0	60.0
500	44.4	3.02

Table.2 shows the results by marginal distribution of 2-D FFT. The number of feature vector is 24, which is twice as $6 \times 6 = 12$ dimensions both for static and dynamic features. It is observed that the marginal distribution is insufficient to represent lip characteristics.

Table.3 is the comparative results of Gaussian mixture and tied-mixture for a various number of training words. This indicates that tied-mixture HMM is robust against the number of training words.

Table.4 shows the results for 100 and 500 words tests using the 70 dimensional feature vector. It is shown that tied-mixture modeling improves visual recognition accuracy by 2-3 % and the method proposed here achieves the performance of 85.0% for 100 word tests. Although it is difficult to compare, this is quite high performance compared to previous studies[1]. It is also advantage that our method is based only on nonparametric features which don't suffer from parameterization errors.

Table 3. Visual speech recognition accuracy vs. number of training words(%)

Training Words		4740	2620	1048
100 Word Test	GM	84.0	81.0	66.0
100 Word Test	TM	85.0	85.0	72.0
500 Word Test	GM	58.2	54.4	44.0
500 Word Test	TM	60.4	56.4	48.6

Table 4. Visual speech recognition accuracy using Gaussian mixture(GM) HMMs and tied-mixture(TM) HMMs(%)

#words	Closed		Open	
	GM	Tied-Mix	GM	Tied-Mix
100	83.0	87.0	84.0	85.0
500	69.4	71.4	58.2	60.4

4. AUDIO-VISUAL INTEGRATION

This paper compares two kinds of integration method such as early integration and late integration[5, 10].

1. Early Integration
   Early integration is based on calculation of likelihood using HMMs which trained by composite vectors of multiple-streams of speech vectors (MFCC + ΔMFCC + ΔPower) and visual vectors (BPF + ΔBPF). Each phoneme has only one HMM trained by composite vectors. Stream weights are changed according to SNR. The weighting is carried out as following,

$$b_{ij}(o_t) = b_{ij}^{audio}(o_t)^{\lambda_{audio}} \times b_{ij}^{visual}(o_t)^{\lambda_{visual}}$$

   Here, $b_{ij}(o_t), b_{ij}^{audio}(o_t), b_{ij}^{visual}(o_t)$ are output probabilities of transition from state i to j for composite vector, audio vector and visual vector, respectively. λ_{audio} and λ_{visual} are weighting coefficients for audio vector and visual vector, respectively, which satisfies,

$$\lambda_{audio} + \lambda_{visual} = 1.$$

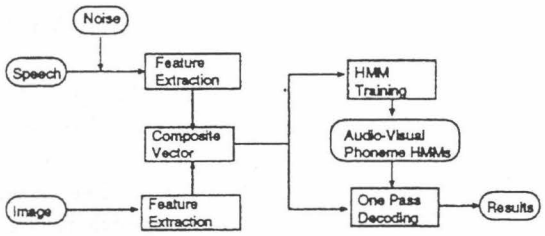

Figure 3. Block diagram of early integration

2. Late Integration
   Late integration is based on calculation of weighted combination of final log likelihoods from a audio HMM and a visual HMM sequences associated to the same word. Each phoneme has two HMMs for speech and visual vectors.

$$P(X|M_i) = P(X_{audio}|M_i^{audio})^{\lambda_{audio}} \times P(X_{visual}|M_i^{visual})^{\lambda_{visual}}$$

   Here,
   $P(X|M_i), P(X_{audio}|M_i^{audio}), P(X_{visual}|M_i^{visual})$ are probabilities of composite vector sequence X from late integration, audio vector sequence

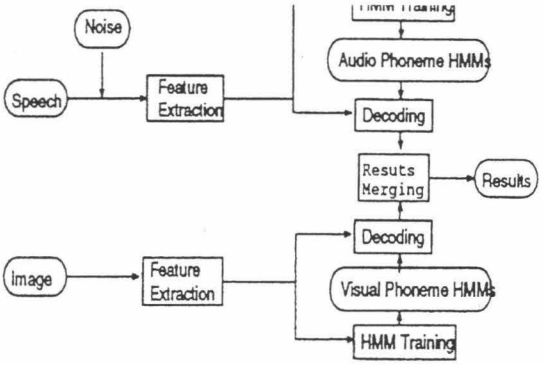

Figure 4. Block diagram of late integration

   X_{audio} from audio HMMs M_i^{audio} and visual vector sequence X_{visual} from visual HMMs M_i^{visual}, respectively. The i is word number. The weighting coefficients also satisfies,

$$\lambda_{audio} + \lambda_{visual} = 1.$$

Audio vector is composed of 33 dimensional feature vectors (16MFCC + 16ΔMFCC + ΔPower). HMMs for speech is 55 tied-mixture monophone HMMs. Number of distributions of tied mixture distribution are 256, 256 and 128 for MFCC, ΔMFCC and ΔPower, respectively. The feature vector for lip image is the same as described in previous section. Each model consists of 5 states including an initial and a final states in early integration. The same number of words is used for audio and visual HMM training. Since frame shifts are different, the same visual vectors are filled for 4 frames to synchronize the frame shift of visual vector to that of speech vector in early integration. In early integration, HMMs are trained using various kinds of stream weights for integration. White Gaussian additive noise is used as an audio noise source. Each word is represented as a phoneme sequence by the network grammar. An input observation vectors are decoded by Viterbi decoding.

Fig.5 and fig.6 show the experiment results. The integration achieves 15% and 5% improvements in SNR=10dB by the early and the late integration compared to visual recognition rates. It is confirmed that the integration improves the recognition accuracy in the range of from -10dB to +20dB by early integration and from +5dB to +20dB by late integration. The improvements in early integration compared to the original performance by either audio or visual information are larger than that by late integration. However, it is shown that the recognition rates obtained from early integration are slightly worse than that of late integration. This slight degradation in early integration is caused by filling the same visual features in order to synchronize the visual frame rate to speech frame rate. The difference between two

kinds of integration is relatively small compared to previous studies. This results suggest that the early integration will able to improve the performance if visual recognition is accurate and that more sophisticated integration needed for early integration.

Fig.7 summarizes the results. The late integration always realizes better performance than either of audio, visual and early integration performance.

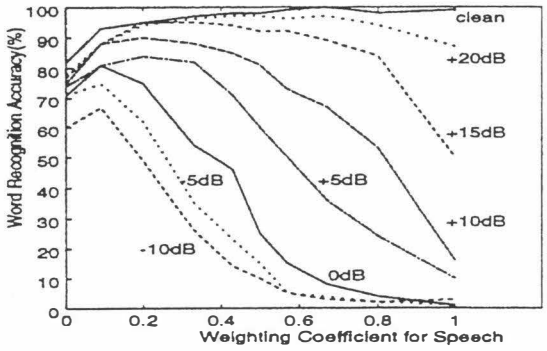

Figure 5. Early Integration

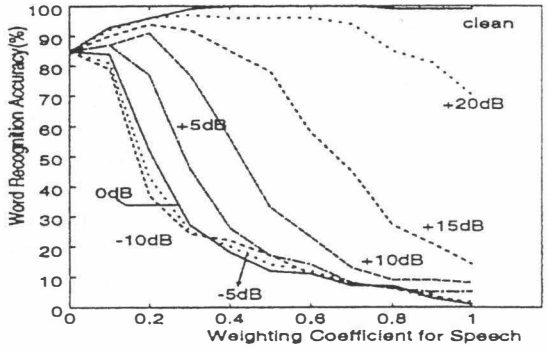

Figure 6. Late Integration

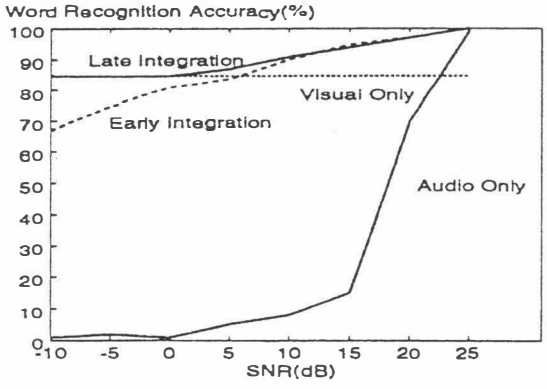

Figure 7. Comparison of Integration Methods

5. CONCLUSION

This paper describes the large speech and image synchronous database, the improved sub-word HMM modeling of lip-image and their integration to improve speech recognition. Speech recognition experiment results show that tied-Mixture HMMs improve lip image recognition accuracy and that the speech and lip image integration improves speech recognition accuracy under various kinds of SNR environments. Two types of audio-visual integration, early integration and late integration are compared. The late integration outperforms the early integration. This result implies independence between audio and visual information. Further study which makes use of correlation between audio and visual information seems to be needed.

REFERENCES

[1] D.G.Stork, M.E.Hennecke, "Speechreading by Humans and Machines", NATO ASI Series, Springer, 1995

[2] E.Petajan, "Automatic Lipreading to Enhance Speech Recognition", Proc.CVPR'85

[3] B.Yuhas, M.Goldstein. Jr,T.Sejnowski, "Integration of Acoustic and Visual Speech Signals Using Neural Networks", IEEE Communications Magazine, pp65-71, 1989

[4] C.Bregler, H.Hild, S.Manke, A.Waibel, "Improving Connected Letter Recognition by Lipreading", Proc.IEEE ICSLP93

[5] A.Adjoudani, C.Benoit, "Audio-Visual Speech Recognition Compared Across Two Architectures", Proc.EUROSPEECH95

[6] P.Silsbee, "Computer Lipreading for Improved Accuracy in Automatic Speech Recognition", IEEE Trans. on Speech and Audio, Vol.4. No.5,1996

[7] D.Chandramohan, P.Silsbee, "A Multiple Deformable Template Approach for Visual Speech Recognition", Proc.ICSLP96

[8] P.Duchnowski, U.Meier, A.Waibel, "See Mee, Hear Me: Integrating Automatic Speech Recognition and Lip-Reading", Proc.ICSLP94

[9] J.Luettin, N.Thacker, S.Beet, "Visual Speech Recognition Using Active Shape Models and Hidden Markov Models", Proc.IEEE ICASSP96

[10] M.Alissali, P.Deleglise, A.Rogozan, "Asynchronous Integration of Visual Information in an Automatic Speech RecognitionSystem", Proc.ICSLP96