DESIGN OF AUDIO-VISUAL INTERFACE FOR AIDING DRIVER'S VOICE COMMANDS IN AUTOMOBILE ENVIRONMENT

Kihyeon Kim[1], Changwon Jeon[1], Junho Park[1], David K. Han[2] and Hanseok Ko[1]

[1] Dept. of Electronics & Computer Engineering, Korea University, Seoul, Korea
[2] Naval Academy, Annapolis, Maryland, USA
khkim@ispl.korea.ac.kr, cwjeon@ispl.korea.ac.kr, jhpark@ispl.korea.ac.kr, han@usna.edu, hsko@korea.ac.kr

ABSTRACT

This paper describes an information-modeling and integration of an embedded audio-visual speech recognition system, aimed at improving speech recognition under adverse automobile noisy environment. In particular, we employ lip-reading as an added feature for enhanced speech recognition. Lip motion feature is extracted by active shape models and the corresponding hidden Markov models are constructed for lip-reading. For realizing efficient hidden Markov models, tied-mixture technique is introduced for both visual and acoustical information. It makes the model structure simple and small while maintaining suitable recognition performance. In decoding process, the audio-visual information is integrated into the state output probabilities of hidden Markov model as multistream features. Each stream is weighted according to the signal-to-noise ratio so that the visual information becomes more dominant under adverse noisy environment of an automobile. Representative experimental results demonstrate that the audio-visual speech recognition system achieves promising performance in adverse noisy condition, making it suitable for embedded devices.

Index Terms— Audio-visual speech recognition, lip-reading, multistream hidden Markov models, tied-mixture models.

1. INTRODUCTION

Employing automatic speech recognition (ASR) system for enabling automobile information services such as voice activated navigation system becomes a formidable challenge due to noisy automobile environment. The problem becomes particularly pronounced when the desired speech is inadvertently contaminated by other passengers' voice, or non-stationary road noises such as horns or sirens.

However, most of acoustical noise suppression approaches perform poorly under interfering noises because the noise component is estimated with the assumption that noise is stationary [1~3].

In this paper, we present the Audio-Visual Speech Interface (AVSI) system extracting driver's audio-visual features as a solution to ensure stable recognition performance in various noisy environments. Specifically, we designed a bimodal information modeling and integration technique which exploit lip reading as an additional feature [4~7].

The design of the AVSI system consists of three modules. The first module handles extracting the driver's visual feature, e.g., lip shape, for visual mapping of the spoken utterances. It is obtained by employing the Active Shape Model (ASM). The second is the acoustic module wherein features are extracted by the conventional method of Mel-Frequency Cepstrum Coefficients (MFCCs) and a priori SNR.is estimated statistically. The third module performs information fusion by weighting the state output probabilities for each stream according to the estimated a priori SNR. Subsequently, lower SNR assigns higher weight on the state output probabilities in visual stream. For decoding process in the fusion module, Tied-Mixture Hidden Markov Model (TMHMM) structure is applied [8]. In this case, states use one global pool of Gaussian mixtures for each information channel. Such information modeling technique provides ability to handle the sparse data problem in training process, simply by reducing the number of Gaussians and thereby dramatically reducing the model complexity. Moreover, it can minimize the computational requirement effectively on calculation of likelihoods when it is combined with relevant ancillary refinements, such as beam control.

We demonstrate through representative experiments that the design is suitable for automobile embedded devices and produces reasonable ASR performance under severe noise conditions as well as under speech contamination by third party utterances.

2. VISUAL FEATURE EXTRACTION

In order to detect lip region, the input image is first converted from the RGB color space to the YCbCr space for its better ability of discerning features for lips. In general, Cr and Cb values of the lip region are higher than those of other regions and Cr values are relatively higher than those of Cb. For Y component, it is higher in the upper part of the lip and is lower at the corners.

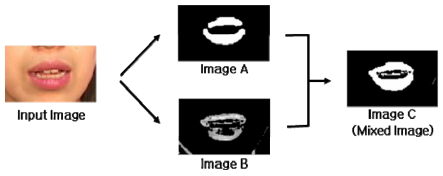

Fig. 1. Segmentation for lip region

In Fig. 1, the image A is the binary image which is extracted by applying a fixed threshold on the 'MouthMap'. The MouthMap, a term for face recognition, is the lip image composed of only chroma components (Cr, Cb).

Image B is generated by the following equation:

$$\text{"Image B"} = (\max(Cr - Y, 0) - \max(Cb - Y, 0))^2 \quad (1)$$

In (1), $\max(Cr-Y,0)$ and $\max(Cb-Y,0)$ eliminate most of the face region except the lip because chroma values are relatively higher in lip region than luminance values. The difference between two terms emphasizes the lip part by considering the relative Cr component to Cb. As a result, the image A represents the middle region of lip well and the image B captures the edges of the lip's upper part and the corners. These two images are mixed to the binary image C for detection of the lip region.

ASM is applied to extract lip points from the detected images as shown in Figure 2. Two models for the open and closed mouth states are made by mean shape models using Principal Component Analysis (PCA). Open mouth model consists of 16 points representing the inside and the outside parts of lip and closed mouth model consists of 10 points for only the outside part of lip as shown in Fig. 2 [9].

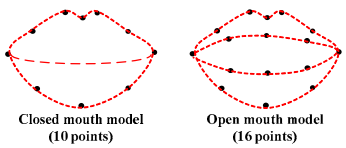

Fig. 2. Mean Shape model (Closed mouth and Open mouth)

ASM forms the lip model equation as

$$X = \bar{X} + \mathrm{Pb} \quad (2)$$

where X is a lip model, $\bar{X}$ is the mean shape model, P is a matrix of eigen vectors by PCA and b is a parameter for the variation of the mouth shape, respectively.

Because the lip image is binary, by counting black pixels inside the detected lip, it can be easily determined whether the mouth state is "open" or "closed." In (2), the appropriate mean shape model is applied according to the mouth state.

In general, ASM is an iterative method to converge to the minimum of the maximum energy defined by equation (4) at each point until the estimated position of the lip model is well aligned to the detected position of the input image. The main criterion is given by

$$\hat{d\mathrm{b}} = \arg\min_{\mathrm{b}+d\mathrm{b}} E(dx, d\mathrm{b}) \quad (3)$$

where E represents the energy function defined by

$$E = (F(dx, d\mathrm{b}))^{\mathrm{T}} W (F(dx, d\mathrm{b})) \quad (4)$$

$$F(dx, d\mathrm{b}) = \left((X + dx) - (\bar{X} + \mathrm{P}(\mathrm{b} + d\mathrm{b}))\right) \quad (5)$$

and W is an arbitrary cost matrix to maximize (4), respectively.

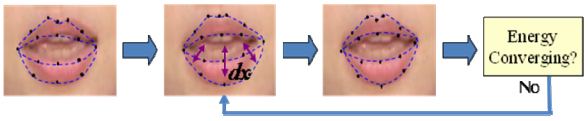

Fig. 3. Tracking lip shape for ASM

For the AVSI system, the visual features are constructed from the detected lip points. The visual stream consists of the width and the height of mouth, and the four selected points that are normalized by mean subtraction to attenuate the effect of individual lip size and to ensure the robustness as shown in Fig. 4.

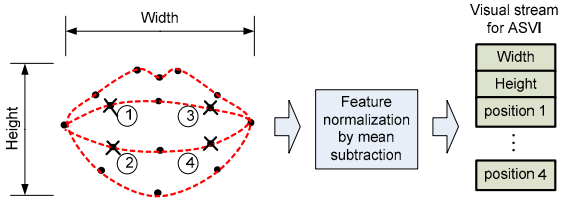

Fig. 4. Construction of visual feature stream for AVSI system

3. SNR DEPENDENT AUDIO-VISUAL INFORMATION COMBINATION

After the visual feature is extracted by the ASM, linear

interpolation on visual feature is employed to synchronize with the acoustic features. These features then form the multistreams for the decoding process. The acoustic and fusion modules are described in the following subsections.

3.1. Acoustic feature extraction and estimation of a priori SNR

In the acoustic module, conventional MFCC method is applied to extract features from the audio signal as shown in Fig. 5.

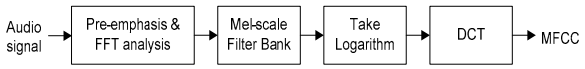

Fig. 5. Schematic diagram of MFCC feature extraction

A priori SNR ξ is estimated by a simple spectral subtraction algorithm as.

$$\xi = \alpha \cdot 10\left[\log 10\left[\frac{1}{L}\sum_{l\in L}\sum_{all\,k}\left[\Phi_Z(k,l)-\Phi_N(k)\right]\right]-\log 10\left[\sum_{all\,k}\Phi_N(k)\right]\right] \quad (6)$$

where k, l are the frequency and frame index; α is the compensation factor; L is the interval where the desired speech signal is activated; Φ_Z, Φ_N are power spectral densities for the input signal and the estimated noise respectively.

The noise power spectrum is estimated by averaging audio signals in the input stream where the desired speech signal is not expected to be present. The factor α is used to compensate the estimated SNR when the noise becomes significant including cases such as non-stationary components like interfering speech or sirens. Thus, when α is set to be lower than one, the AVSI system would place more weights on the visual stream. It has been observed that the estimation procedure does not need to define exact values of α. Instead, discretely assigned values are often sufficient for the weights. Table 1 shows the apparent optimal audio-visual weights to ensure successful recognition performance of the AVSI system.

Table. 1. Optimal audio-visual weights according to various SNR levels

SNR (dB)	SNR<0	0<SNR<5	5<SNR<10	SNR>10
A-V* weights	(0, 1)	(0.6, 0.4)	(0.8, 0.2)	(0.9, 0.1)

(* A-V : audio–visual pair)

In an adverse acoustic noise condition, the visual weight becomes higher than the acoustic one since the visual information is relatively more reliable than the acoustic information.

3.2. Audio-visual model combination

Generally, three types of structures are investigated to integrate audio-visual information: early integration, late integration and hybrid integration [5]. However, most of the recent work has focused on the hybrid integration method due to its flexibility and robustness. This paper develops the AVSI system based on the hybrid integration technique which is also known as multistream [4,6]. We introduce the TMHMM as the model structure suitable for embedded applications such as telematics [8]. Fig. 6 describes the overall process of the proposed multistream integration.

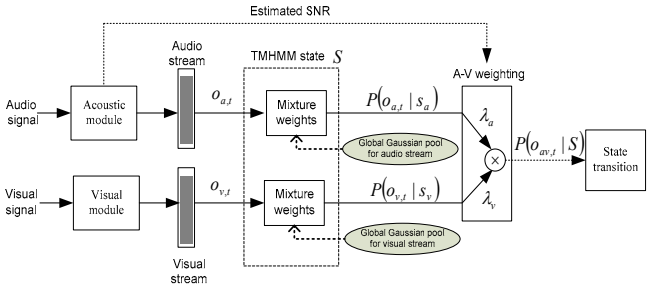

Fig. 6. Schematic diagram of TMHMM based multistream audio-visual integration

In the fusion process, state output probability of HMM is given by

$$P(o_t \mid S) = P\left(o_{a,t} \mid s_a\right)^{\lambda_a} P\left(o_{v,t} \mid s_v\right)^{\lambda_v} \quad (7)$$

where $o_{a,t}, o_{v,t}$ are audio and visual feature streams at a frame t; S is the TMHMM state which is composed of substates s_a for the audio stream and s_v for the visual stream; λ_a, λ_v are weighting factors that are constrained by relations $\lambda_a, \lambda_v > 0$ and $\lambda_a + \lambda_v = 1$ respectively. That is, the state output probability is the linear combination of emission probabilities from corresponding substates in log-domain.

This audio-visual fusion technique has several advantages in training and decoding. First, it requires the searching process only once and thus would reduce computational requirements significantly. This is due to the fact that the emission probabilities are combined within one state before the state transition occurs, and therefore it prevents any parallel search. Although several previous work had shown that the audio-visual state propagation scheme [5] can enhance the recognition performance, improvements obtained were not significant enough for the required increase in the computational load in decoding. The second advantage is that the audio and the visual models can be trained independently if the state transition probability relies only on the acoustic model. As mentioned earlier, the state output probability is composed of simple audio and visual

emission probabilities. In other words, the substates s_a and s_v do not effect each other in decoding and can be trained independently. Moreover, the introduction of TMHMM makes the model structure simple by having to construct just one global Gaussian pool for each information channel. In this way, every state holds the pools in common and has only mixture weights. Hence, for training, the data sparse problem can be handled simply by decreasing the number of Gaussian components in the pool while the computational load for the decoding can be reduced by the beam pruning according to the mixture weights. Finally, weights for audio and visual stream are determined automatically by the estimated SNR in section 3.1. Therefore, the system can adapt to the variation of the environmental noise.

These advantages ensure that the audio-visual speech recognition system has reasonable performances and can be easily implemented for automobile embedded devices.

4. EXPERIMENTAL RESULTS

For experiments, acoustic models were synthesized by using a decision tree over previously well trained TMHMM's. Visual models were trained by real video data and formed a global Gaussian pool. State transition probabilities rely on the acoustic model. As test data, we used the audio-visual DB which was composed of recording by twenty Korean speakers. Each speaker pronounced twenty arbitrarily selected Korean words twice. The total number of test files was 800. Automobile noises including the third utterance were added to the clean audio data with various ranges of SNR from -10dB to 20dB.

Fig. 7 shows Word Error Rates (WERs) of the proposed audio-visual speech recognition system in comparison with the baseline recognizer using audio information only.

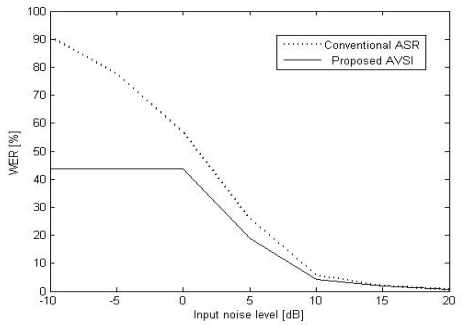

Fig 7. WER of the proposed audio-visual speech recognition system in comparison with conventional ASR system.

The conventional ASR system naturally has higher WERs at lower SNR values. The performance falls rapidly below 10dB of SNR. On the other hand, the proposed audio-visual speech interface system ensures the recognition performance under 43% of WER even when the input SNR is lower than 0dB as well as gives improved results constantly in overall SNR range in comparison with the conventional system.

5. CONCLUSION

This paper describes the implementation of an audio-visual speech recognition system designed to ensure the performance in adverse noise environment. The system uses mustistream and TMHMM structure suitable for automobile embedded devices. The approach allows the bimodal fusion to be simple while maintaining deployable level of recognition performance. Moreover, the proposed SNR estimation and the automatic weighting technique enable the system to actively adapt to various noise conditions. Several experiments have shown that the AVSI system can be a reliable solution, compared to the conventional ASR system, even in noisy environment including interfering noise such as third-party speech.

ACKNOWLEDGEMENT

This research was supported by the MIC (Ministry of Information and Communication), Korea, Under the ITFSIP (IT Foreign Specialist Inviting Program) supervised by the IITA (Institute of Information Technology Advancement)

REFERENCES

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-27, pp. 113-120, Apr. 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, pp. 443-445, Apr. 1985.

[3] S. Gannot *et al*., "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, pp. 1614-2001, Aug. 2001.

[4] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, Vol. 2, pp. 141-151, Sept. 2000.

[5] G. Potamianos *et al*., "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of The IEEE.*, vol. 91, pp. 1306-1326, Sept. 2003.

[6] J. F. Perez *et al*., "Lip reading for robust speech recognition on embedded devices," *IEEE International Conference on Acoustics, Speech and Signal Processing, 2005.*

[7] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 1082-1089, May. 2006.

[8] J. Park and H. Ko, "Achieving a reliable compact acoustic model for embedded speech recognition system with high confusion frequency model handling," *Speech Communication*, vol. 48, pp. 737-745, June. 2006.

[9] A. Caplier, "Lip Detection and Tracking," *11th International Conference on Image Analysis and Processing*, pp. 8-13, 2001.