한국음향학회지 제28권 제8호 pp. 832~842 (2009)

A Novel Integration Scheme for Audio Visual Speech Recognition

Than Trung Pham*, Jin Young Kim*, Seung You Na*

*School of Electronics & Computer Engineering Chonnam National University
(접수일자: 2009년 11월 13일; 채택일자: 2009년 11월 17일)

Automatic speech recognition (ASR) has been successfully applied to many real human computer interaction (HCI) applications; however, its performance tends to be significantly decreased under noisy environments. The invention of audio visual speech recognition (AVSR) using an acoustic signal and lip motion has recently attracted more attention due to its noise-robustness characteristic. In this paper, we describe our novel integration scheme for AVSR based on a late integration approach. Firstly, we introduce the robust reliability measurement for audio and visual modalities using model based information and signal based information. The model based sources measure the confusability of vocabulary while the signal is used to estimate the noise level. Secondly, the output probabilities of audio and visual speech recognizers are normalized respectively before applying the final integration step using normalized output space and estimated weights. We evaluate the performance of our proposed method via Korean isolated word recognition system. The experimental results demonstrate the effectiveness and feasibility of our proposed system compared to the conventional systems.

Keywords: Audio Visual Speech Recognition, Reliability, Late Integration, Hidden Markov Model

ASK subject classification: Speech Signal Processing (2)

I. Introduction

Automatic speech recognition (ASR) has attracted significant interest of many researchers around the world due to its exciting applications and challenges. A successful ARS can be applied to well-defined applications like dictation and medium vocabulary transaction processing tasks in controlled environments. However, the performance of ASR systems has not yet reached the required level for real applications. This is because the accuracy of ASR systems will be degraded rapidly in real applications because of several factors such as channel, environment, *etc.*. The most important one is the effects of noisy environments. When speech is taken in noisy conditions such as bus, babble, car, etc., its acoustic modality is far different from that of training data.

Corresponding author: Jin Young Kim (beyindi@jnu.ac.kr)
School of Electronics & Computer Engineering, chonnam National University, 300 Yongbong-Dong Buk-Gu, Gwangju 500-757, Korea

However, human speech perception considers not only an acoustic signal but also lip motion. Speech is more intelligible if the face of the speaker can be seen, especially in noisy conditions. Thus, more robust automatic speech recognition is possible if visual information can be integrated with traditional acoustic systems. There have been many publications [1-5] against integrated audio visual approach for enhancement of the performance of speech recognition systems. They claimed that the integration of audio and visual modalities significantly improves the accuracy, especially in noisy conditions. Although there are a variety of proposed approaches to audio visual fusion, they can be divided into two categories: early integration and late integration. The late integration method is proved to be more effective than the early one by many experimental results.

Generally, the late integration based AVSR systems work by the following steps. First, the acoustic and visual speeches are recorded and potential features are extracted. Then, these features are inputted

to audio visual recognizers, for example HMMs, respectively to output the corresponding probabilities. Finally, the two probabilities of audio and visual streams are integrated with the use of stream weights or reliabilities. The reliability is used to measure how much each modality contributes to integration. A common way to estimate the reliability is based on the noise level (SNR) [6-8] or voicing index [9] or frame-dependent SNR of audio signals [10]. However, this approach has a problem that the reliability of visual speech is not taken into account. There are other proposed methods to measure the confidence of audio visual modality based on the output probabilities of the corresponding HMMs, called entropy based confidence as in [11]; in [12] the authors use a further neural networks step to determine the weight of each modality based on score-based reliability.

In this paper, we introduce a more robust reliability measurement applied to the audio visual late integration. Our proposed method combines the model based reliability and the signal based reliability. The model based reliability is calculated by using HMMs of audio and visual models. The motivation of this calculation is to measure the confusability of a given vocabulary of ASR which is one of the factors causing the degradation of the performance. The signal based reliability is based on noise information of audio signals including a SNR level and a noise type. These two reliabilities are fused via a mapping function optimized by particle swarm optimization (PSO) [13] to compute the weights of audio and visual modalities. At the integration step, the output probabilities of audio and visual recognizers are normalized to the range [0 1] respectively, and then summed up with the appropriate estimated weights.

II. Audio Visual Model based Reliability

In speech recognition, even though the speech signal is perfectly clean, the performance still cannot reach to 100 %. This is because there are words in the vocabulary being similar or confusable. For example, 'presents' is easily misunderstood as 'presence'; 'might' can be misrecognized as 'night' sometimes. This leads to the confusability definition in the speech recognition system. Confusable word problems are also totally correct under the visual speech recognition system. As we know, the performance of visual speech recognition is always lower than the audio case because visual information or lip motion is distinguished less than audio signals. It is true that human conversation is difficult to understand if we just look at the speaker's mouth.

2.1. Word Confusability and HMM Distance

In the speech recognition system using HMM, speech is modeled by HMM, and it is often to measure how confusable the two words are. This comes to the calculation of the distance between two HMMs or HMM metric. The common approach is computing the Bayes error of two HMMs. However, there are no analytically and numerically tractable closed form methods to solve this problem so far. A common alternative, that is numerically approachable, is HMM divergence instead of HMM metric. The two most widely used methods are based on Kullback Leibler [14] and Bhattacharyya [15] divergence.

Confusability of two words can be defined as the similarity or distance of two HMMs [16]. The smaller the distance is, the more confusing the two words are. The common method to calculate the HMMs distance is based on the Cartesian product of two HMMs (Figure 1) resulting in the state-based weighted edit graph. The weight for each edge of this graph is the distance of two GMMs appropriate with two states of two words respectively. The distance of two GMMs can be estimated by Kullback Leibler or Bhattacharyya divergence [17-19]. The final HMM distance is then the total of weights in the shortest path of the graph [20][21] (Equation 1).

$$D = \sum weights\ of\ shortest\ path \quad (1)$$

In our study, we used Bhattacharyya divergence which is proved to be better than others methods [18] to compute word confusability. By using word confusability, we are able to calculate the predictable confusion matrix of a specific word group in which an element of a confusion matrix is the distance of two corresponding HMMs. Confusion matrix is commonly used in predicting phonemics confusions in speech recognition systems. However, the idea that we used confusion matrix to estimate the model based reliability of a word group is comparatively novel and will be interpreted in the next sections.

2.2. Normalized HMMs Distance

In a specific ASR system, different words in a vocabulary can be modeled by the different HMMs structure with various numbers of states. If two words have a larger number of states, the weighted edit graph will be bigger; therefore the total HMM distance or the shortest path is greater than the two words having a smaller number of states. This leads to an incorrect confusion matrix. In order to solve this problem, we need to normalize the HMM distance associated with the number of states of HMMs. In our study, we proposed a simple normalization algorithm which is done by dividing the summation of weights of the shortest path by the number of steps or edges of the shortest path. By doing this, the total distance will not increase as long as the size of the graph that is proportional to the number of states of HMMs.

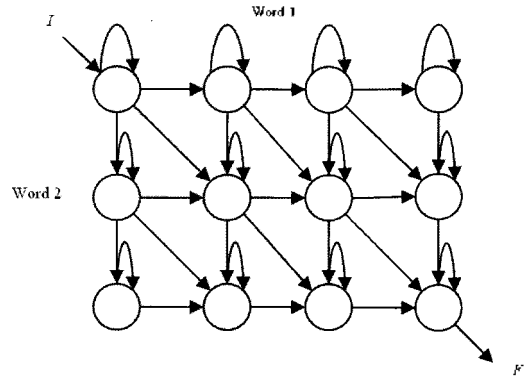


Figure. 1. The Cartesian product of two HMMs. The first HMM has 4 states; the second one has 3 states. The weight of edges can be computed underlying the pair of GMMs.

$$D_{Normalized} = \frac{\Sigma weights\ of\ shortest\ path}{size\ of\ shortest\ path} \quad (2)$$

2.3. Confusability of a Word Group

Confusability of a given word group is defined as a measurement that enables to represent the complexity of a word group with the point of recognition performance. A word group with high confusability will have low performance, and low confusable word groups, similarly, often get high recognition rates. Consequently, we can considerably estimate the model based reliability of a word group based on its confusability. The question is how to calculate the confusability of a given word group? Our idea is using the confusion matrix obtained from HMM distance. Let us assume that we have a group of N words. Each word is modeled by a HMM. Then we can easily compute the N by N confusion matrix for this group using HMMs divergence. In the confusion matrix, each element is the distance of two HMMs which is reverse proportional to the confusability of those two words. The first task in our method is to detect confusable candidates of each word in the group because the couples having less confusability are not useful at all in contributing to the confusability of the group. In our study, we observed that the elements in the confusion matrix having the value less than the mean of the confusion matrix are usually confusable candidates. This means that the two words associated with the column or row of those elements could be confusable words. Furthermore, for each word in a group, the rest of (N-1) words compete each other to gain the confusable position instead of cooperation to total confusability of that word. For example, in a group, word A can be confusable with two other words, and word B can be confusable with only one. Then, it is not always true that confusability of A is greater than B. In fact, it depends on how much confusability of word A or B has with others. Based on those properties, the proposed algorithm for calculating confusability can be described as below.

- *Given confusion matrix* $C_{N\times N}$
- *Compute threshold: th = mean(C)*
- *For each row i of confusion matrix C*
  - *o Find elements in row i except the diagonal one that smaller than th*

$$E=\{e\,|\,C(i,e)<th, e!=i\} \quad (3)$$

  - *o Compute confusability of word I:*

$$C_i = e^{-mean(C(i,E))} \quad (4)$$

- *Compute confusability of group*

$$C=\frac{\sum C_i}{N} \quad (5)$$

The output value of this algorithm belonging to the range (0, 1) represents the confusability of a group.

Proof:

$\forall\, i, j \in [1, N], C(i, j) > 0,$

$then, mean(C(i, E)) > 0$

$therefore\ 0 < C_i = e^{-mean(C(i,E))} < 1$

$Finally, we\ have\ 0 < C = \frac{\sum C_i}{N} < 1$ ■

2.4. Confusability based Audio and Visual Reliability

It is clear that word confusability can cause the decrease of speech recognition performance (see Figure 2). Thus confusability can be considered as confidence of audio and visual modalities. However, this confidence is based on the models rather than the signals. So we can compute it offline before the real online testing step. In this part, we will discuss our experiment set-up and describe how to map confusability of audio visual models to audio visual confidence.

In our study, we set up 100 word groups, 10 words for each group. We use clean data in order to avoid the case that a low SNR signal can affect the performance of recognition. The confusability of both audio and visual models is computed for all word groups. In order to introduce the relation between audio visual confusability and audio visual reliability, we use the particle swarm optimization (PSO) method to estimate the weights of modalities for each group in the case of late audio visual integration $P'_{av} = \alpha P'_a + (1-\alpha) P'_v$ (α is weight of audio speech for a group) that gives the best performance. This means that each word group has an optimized weight of audio modality and weight of visual one.

Physically, the weight of audio speech tends to be larger than the weight of visual speech if the confusability of audio is smaller than that of visual. On the other hand, the contribution of audio to final integration will increase much if visual confusability is large and audio confusability is small. This shows that the audio visual confidence will be proportional to the ratio of audio and visual confusability. In fact, Figure 3 shows this relation. The vertical axis is the optimized weights of audio modality, and the horizontal axis is the ratio of logarithm of audio visual confusability. The logarithm of confusability can act as the contribution. The confidence of audio and

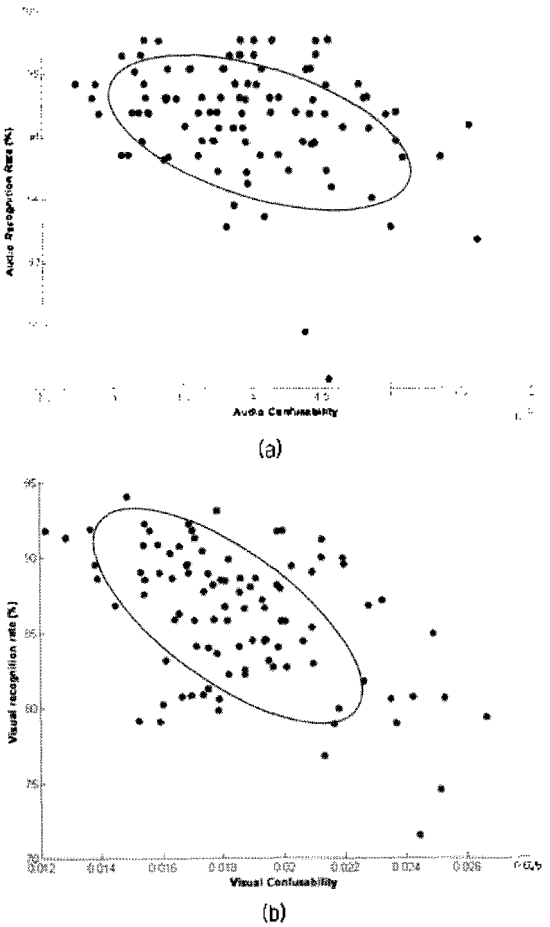


Figure. 2. Distribution of confusability of 100 groups, 10 words for each and associated recognition rate: a: audio case; b: visual case.

visual modalities for a word group can be estimated as:

$$\Upsilon_A = 1.1\frac{\log(C_A)}{\log(C_V)} - 0.64 \quad (6)$$

$$\Upsilon_V = 1 - \Upsilon_A \quad (7)$$

where C_A and C_V are confusability of audio and visual word group, respectively.

III. Audio Visual Signal based Confidence

In the speech recognition system, the quality of signals plays a significant role in the recognition rate. The audio speech is usually affected by background noise while the visual speech is affected by illumination change or contains just a partial lip region. For the audio visual speech recognition system, if the signal quality is high, its contribution to final audio visual integration should be large and vice versa. In real applications, the audio speech and the visual speech are varied; therefore the contribution of them to the combination step should be adaptive. For example, if the environment is noisy with low SNR, the visual speech should be the main stream in the integration step. If the lighting condition is bad, then the audio speech in turn is the main stream. However, in this study, we assume that the speaker talks under good lighting conditions; there is not much illumination change; and the lip is available in visual stream. We focus on considering the case of noisy environments because noise exists in most of the real applications. In this section, we introduce the algorithm to estimate the confidence of audio signal based on acoustic noises.

There are many research reports showing that the performance of ASR system is proportional to the SNR value of audio signals. The reliability of audio signals can be estimated directly by using the SNR value. However, in real applications, there are many types of noises, and each noise affects signals differently because different noises have different spectrums. In this section, we introduce the mapping functions that can estimate reliability of audio signals according to different noise types (Figure 4).

In this section, we describe the procedure of how to estimate reliability of audio signals associated with 5 noise types: white, pink, babble, car (volvo), and factory noise. We use NOISEX-92 [22] database for our experiments. Noise samples are added to clean speech signals at different SNR levels from −30 dB to 30 dB to create testing data. The audio only speech recognition is performed on this data. In order to observe the exact influence of noise in speech recognition, we select the word group having little

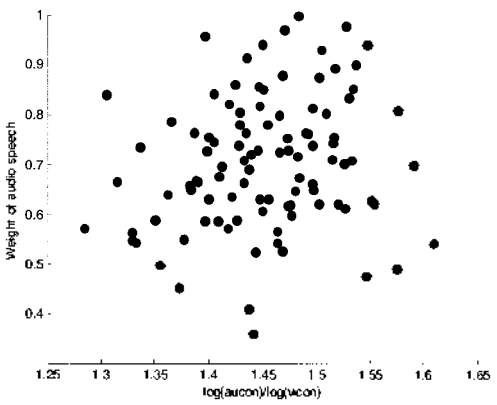


Figure. 3. Distribution of audio reliability and ratio of audio visual logarithm confusability.

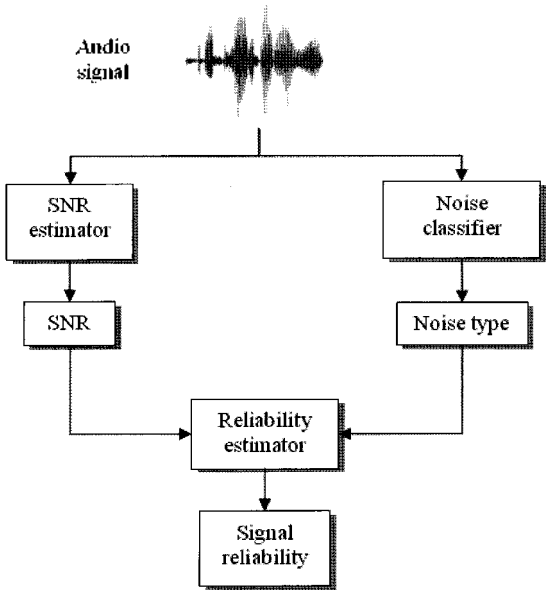


Figure. 4. Diagram of estimation audio signal reliabilityalgorithm.

confusability. Figure 5a shows the distribution of recognition rates associated with different noise types at different SNR levels with the shape of the sigmoid function. The results show that different noises degrade the performance of system differently. White noise influences most while car noise affects less than others. Truly, the recognition rate can tell us how reliable of audio signals is; this shows the close relationship of them. The reliability of speech signal could reach the maximum if the recognition rate is higher than 95 percentages. Similarly, the reliability of speech is getting lower when the recognition rate degrades. Therefore, sigmoid functions (Equation 8) of reliability which is associated with recognition rate, denoted as ε_A, and SNR for the different noise type can be approximated readily based on these experimental observations (Figure 5).

$$\varepsilon_A = \frac{1}{1+e^{-\frac{SNR - m}{s}}} \tag{8}$$

IV. Proposed Integrated AVSR System

In this integration scheme, the confusability and the noise causing the decrease of the performance of ASR system are used to measure the confidences of audio visual modalities. Figure 6 shows our proposed

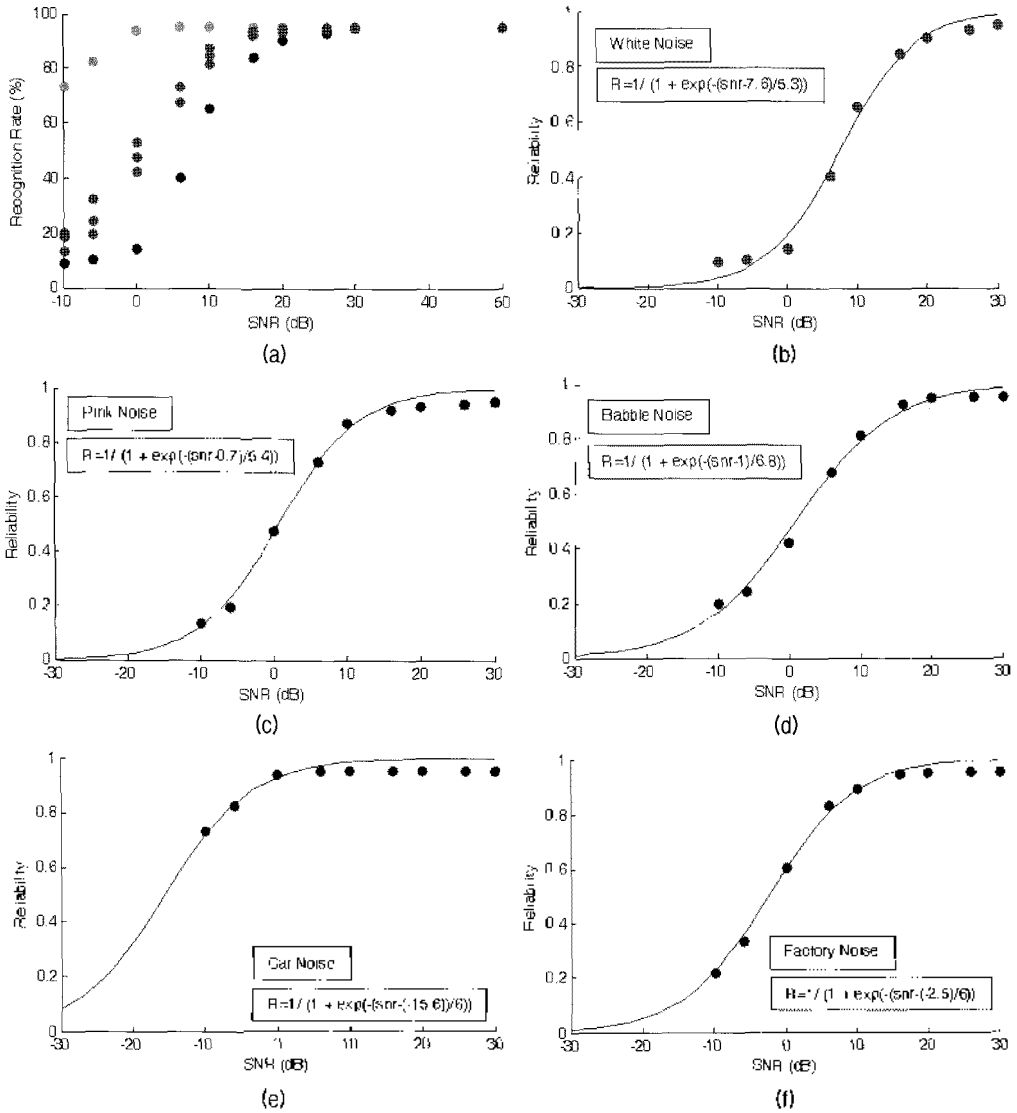


Figure. 5. (a) distribution of recognition rate and SNR value associated with noise types: (b, c, d, e, f) approximated sigmoid function of SNR and reliability of speech signal corresponding white, pink, babble, car and factory noise.

integration scheme for AVSR. In the offline stage, we calculate the confusability of audio and visual word group by using trained HMMs (section 2); then the model based reliabilities of audio and visual are computed. In the online (testing) state, the short portion of the beginning of the recorded audio signal, considered as the noise background, is extracted to be used to estimate the SNR and the noise type; then the signal based reliabilities are estimated (section 3). After that, the final audio and visual reliabilities are estimated by using two previous reliabilities through equation $\lambda = \varepsilon^{\beta}\Upsilon$, where λ is final reliability; ε is signal based reliability; Υ is model based reliability; β is controlling parameter. The controlling parameter helps balance the signal based reliability and the model based reliability because the influences of confusability and noise on the performance of ARS are not the same. Instead, noise degrades the accuracy of AVSR system more significantly than confusability. The controlling parameter is varies with the different noise type. In our study, this parameter is estimated using a PSO algorithm. Table 1 shows various controlling parameters associated with different noise types. In case of visual speech, we assume that there is not much illumination change; so visual signal based reliability is 1. Thus the final visual reliability equals to visual models based reliability. The weights of audio and visual modalities used in late integration are calculated as Equations 11 and 12. At the integration step, the output probabilities of corresponding audio visual recognizers are normalized to the range [0 1] respectively. By doing this, we can avoid the difference of audio and visual probability spaces due to the difference of HMM structure for modeling audio and visual modalities. The normalization works similarly to the score-based reliability. The integration is performed by Equation 13.

- Final reliabilities of audio and visual modalities:

$$\lambda_A = \varepsilon_A^{\beta}\Upsilon_A \tag{9}$$

$$\lambda_V = \Upsilon_V \tag{10}$$

Table 1. Controlling parameter beta associated with various noise types.

Noise	White	Pink	Babble	Car	Factory
Beta	4.74	4.66	5.14	5.35	3.3

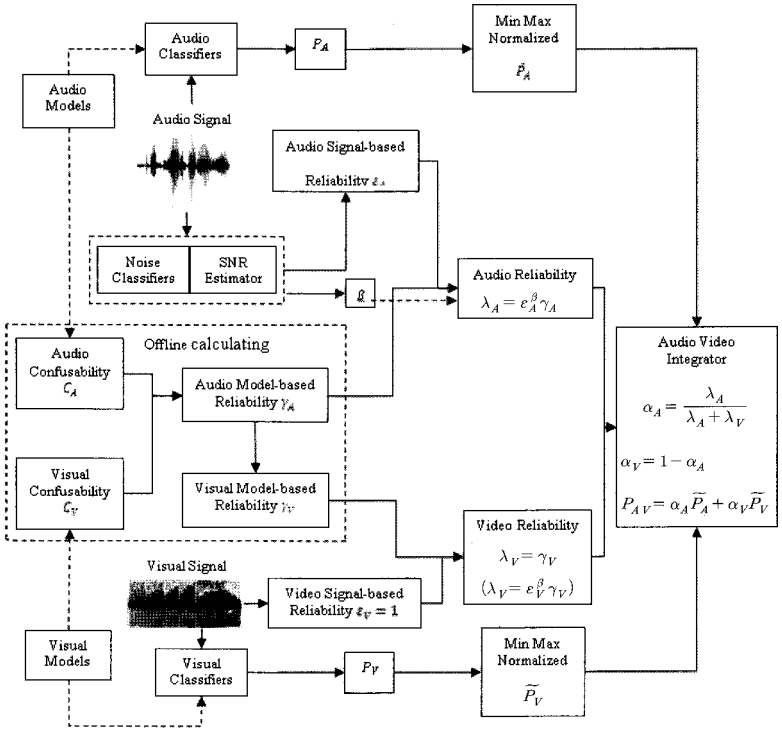


Figure. 6. Proposed audio visual integration.

• Weight of audio and visual modalities:

$$\alpha_A = \frac{\lambda_A}{\lambda_A + \lambda_V} \tag{11}$$

$$\alpha_V = 1 - \alpha_A \tag{12}$$

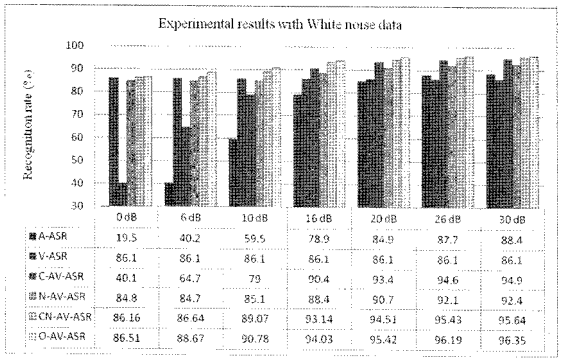
• Audio visual integration

$$P'_{av} = \alpha P'_a + (1-\alpha) P'_v \tag{13}$$

V. Experimental Results

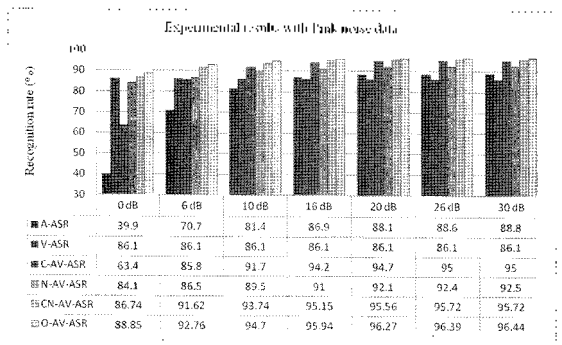
5.1. Database Description

In this study, we use Samsung AVSR database for all experiments. This database is constructed for the purpose of developing AVSR system in mobile phone environment. Data contain 11550 video files recorded from 105 speakers speaking 110 Korean words under three environments: standard (clean), indoor and outdoor. In addition, NOISEX-92 is used to train statistical models for noise classification and add to standard Samsung AVSR database to construct data in various noise conditions. Noises are added to clean data at some specific SNR levels ranging from -30dB to 30dB. We organize the speech recognition vocabulary including 100 groups in which each group contains 10 different words. For each word among 110 words, data from speaker 1 to 80 is used to train HMM, and the rest of data is used for testing.



	0 dB	6 dB	10 dB	16 dB	20 dB	26 dB	30 dB
A-ASR	19.5	40.2	59.5	78.9	84.9	87.7	88.4
V-ASR	86.1	86.1	86.1	86.1	86.1	86.1	86.1
C-AV-ASR	40.1	64.7	79	90.4	93.4	94.6	94.9
N-AV-ASR	84.8	84.7	85.1	88.4	90.7	92.1	92.4
CN-AV-ASR	86.16	86.64	89.07	93.14	94.51	95.43	95.64
O-AV-ASR	86.51	88.67	90.78	94.03	95.42	96.19	96.35

Figure. 7. Comparable results of different recognition approaches using White noise at different SNR.



	0 dB	6 dB	10 dB	16 dB	20 dB	26 dB	30 dB
A-ASR	39.9	70.7	81.4	86.9	88.1	88.6	88.8
V-ASR	86.1	86.1	86.1	86.1	86.1	86.1	86.1
C-AV-ASR	63.4	85.8	91.7	94.2	94.7	95	95
N-AV-ASR	84.1	86.5	89.5	91	92.1	92.4	92.5
CN-AV-ASR	86.74	91.62	93.74	95.15	95.56	95.72	95.72
O-AV-ASR	88.85	92.76	94.7	95.94	96.27	96.39	96.44

Figure. 8. Comparable results of different recognition approaches using Pink noise at different SNR.

5.2. Performance Evaluation

We demonstrate the performance of proposed combined audio visual speech recognition in comparison with various conventional approaches such as visual only ASR (V-ASR), audio only ASR (A-ASR), confusability based integrated audio visual ASR (C-AV-ASR), noise based integrated visual audio ASR (N-AV-ASR), and confusability and noise based integrated audio visual ASR (CN-AV-ASR). The target performance is that of optimal audio visual ASR (O-AV-ASR). The optimal performances are obtained experimentally by using a stochastic optimization technique, PSO, with the variable of audio weighting value α_A. Figures from 7 to 12 show the comparable results of different recognition methods using various noise data at distinct SNR levels. It can be seen from the experimental results that at low SNR levels, N-AV ASR method is better than C-AV-ASR, and at high SNR levels C-AV-ASR is better. At high SNR levels, the audio signal is nearly clean: the audio signal based reliability is reaching to 1; thus the performance of N-AV-ASR is not much enhanced compared to A-ASR. However, C AV-ASR works better because it takes confusability into integration. At low SNR levels, the performance of A-ASR

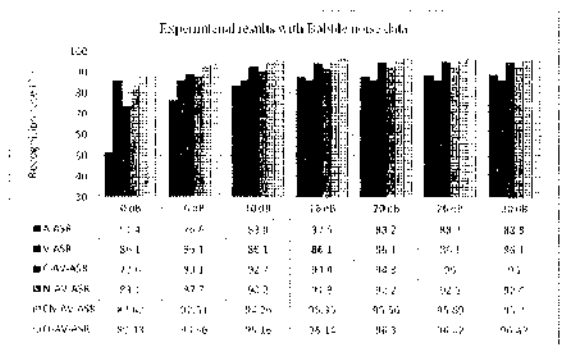


Figure. 9. Comparable results of different recognition approaches using Babble noise at different SNR.

decreases rapidly; the influence of noise on recognition accuracy is more than that of confusability; thus N-AV-ASR works better. The results also demonstrate that the performance of the proposed CN-AV-ASR is always better than N-AV-ASR and C-AV-ASR at any SNR levels, especially nearly reaches the optimal target, O-AV-ASR. Therefore, the performance of our proposed integration approach outperforms the conventional speech recognition methods at any levels of SNR.

VI. Conclusion and Future Direction

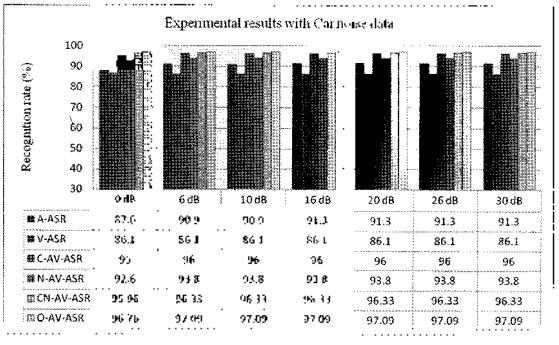
6.1. Conclusion

In this study, we have proposed and constructed a novel scheme for audio visual integration for AVSR. To enhance the overall performance of ASR, we introduced a robust audio visual reliability measurement that involves three factors: SNR level, noise type and confusability. The reliability of audio or visual stream is very important in the late integration because it is used to control the weight of each stream or determine how much each stream contributes to a final decision.

The first factor causing the degradation of recognition is the noise of audio signals or SNR levels. The SNR level is commonly used to estimate the weight of audio stream for audio visual integration. In real applications, there are various noise types such as car, bus, babble etc.; each noise type affects the audio signal differently. For example, the car noise rarely influences the recognition performance although the audio signal has low SNR. Thus, for each type of noise, we introduce a mapping function from the SNR level to reliability of audio signals. This signal based reliability will be one part of audio stream weight for integration.

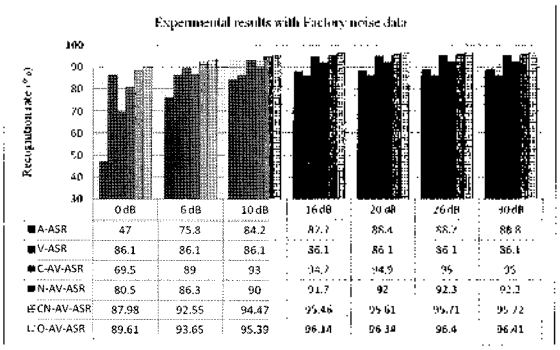
Another factor causing the accuracy degradation of the speech recognition system is the confusability that defines how similar two words are. The confusable words are easily misrecognized even though the signal is clean. When constructing the vocabulary for isolated speech recognition, we are able to calculate the confusability of a word group based on the HMM distances for audio and visual respectively, called model based reliabilities. If the audio word group is more confusable than the visual word group, the audio weight should be less than the visual, and vice versa. In our study, we introduced a mapping function of two variables, audio word group confusability and visual word group confusability; the output of this function is reliability of audio models that will be involved in final reliability of audio stream for integration.

The combination of model based reliability and signal based reliability creates a robust weight measurement for audio visual integration. The model based reliability is demonstrated to be very useful to enhance the performance of AVSR at high SNR levels, while the signal based reliability helps improve the accuracy of ASR when the SNR level is low. This point



	0dB	6dB	10dB	16dB	20dB	26dB	30dB
A-ASR	87.6	90.9	90.9	91.3	91.3	91.3	91.3
V-ASR	86.1	86.1	86.1	86.1	86.1	86.1	86.1
C-AV-ASR	95	96	96	96	96	96	96
N-AV-ASR	92.6	93.8	93.8	93.8	93.8	93.8	93.8
CN-AV-ASR	95.96	96.33	96.33	96.33	96.33	96.33	96.33
O-AV-ASR	96.76	97.09	97.09	97.09	97.09	97.09	97.09

Figure. 10. Comparable results of different recognition approaches using Car noise at different SNR.



	0dB	6dB	10dB	16dB	20dB	26dB	30dB
A-ASR	47	75.8	84.2	87.7	88.4	88.7	88.8
V-ASR	86.1	86.1	86.1	86.1	86.1	86.1	86.1
C-AV-ASR	69.5	89	93	94.7	94.9	95	95
N-AV-ASR	80.5	86.3	90	91.7	92	92.3	92.3
CN-AV-ASR	87.98	92.55	94.47	95.46	95.61	95.71	95.72
O-AV-ASR	89.61	93.65	95.39	96.14	96.38	96.4	96.41

Figure. 11. Comparable results of different recognition approaches using Factory noise at different SNR.

is outstanding compared to conventional audio visual ASR that uses only the noise level for computing the weights of audio and visual streams. Based on this scheme of reliability measurement, we constructed the audio visual integration approach based on late integration. The performance of our proposed scheme has been evaluated via speaker independent isolated word speech recognition. The experimental results demonstrate the effectiveness and the feasibility of our proposed approach compared to the conventional methods.

6.2. Future Direction

In our further study, we will consider the visual signal based reliability for visual audio integration. This should be useful when the illumination changes during the speaker's talks. When the lighting condition is not as good as in testing data, the performance of visual speech recognition will decrease; thus the overall performance of audio visual ASR also degrades. The reliability of visual signals is also measured by the existence of lip in the visual stream. This could happen when the application cannot capture the lip region fully or partially. It is another factor causing the bad performance of visual speech recognition and combined audio visual systems. Therefore the problem of measuring the confidence of visual signals needs to be addressed so that a robust audio visual ASR can be realized for real practical applications.

Acknowledgement

This work was supported by National Research Foundation of Korea Grant funded by the Korean Government (2009-0077345).

References

1. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition," *EURASIP Journal on Applied Signal Processing*, vol. 1, pp. 1274 1288, 2002.
2. Petajan, E.D., "Automatic Lipreading to Enhance Speech Recognition," *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 40-47, 1985.
3. T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9-21, 2001.
4. P. Duchnowski, U. Meier, A. Waibel, "See Me, Hear Me: Intergrating Automatic Speech Recognition and Lipreading", *Proceedings of ICSLP*, pp. 547-550, 1994.
5. G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview," in Issues *in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), MIT Press, Boston, 2004.
6. F. Berthommier, H. Glotin, " A new SNR-feature mapping for robust multistream speech recognition," *Proceedings of International Congress on Phonetic Sciences (ICPhS)*, vol. 1, pp. 711-715, San Francisco, 1999.
7. Md. J. Alam, Md. F. Chowdhury, Md. F. Alam, "Comparative Study of A Priori Signal-To Noise Ratio (SNR) Estimation Approaches for Speech Enhancement", *Journal of Electrical & Electronics Engineering*, vol. 9, no. 1, pp. 809-817, 2009.
8. A. Rogozan, P. Del'eglise, and M. Alissali, "Adaptive determination of audio and visual weights for automatic speech recognition," *Proceedings of European Tutorial Workshop on Audio-Visual Speech Processing (AVSP)*, pp. 61 - 64, 1997.
9. H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," *Proceedings of IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 173 - 176, 2001.
10. M. Heckmann, F. Berthommier, and K. Kroschel, "Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 1, pp. 1260 - 1273, 2002.
11. M. Gurban and J.-Ph. Thiran, " Using Entropy as a Stream Reliability Estimate for Audio-Visual Speech Recognition," *Proceedings of 16th European Signal Processing Conference*, Lausanne, Switzerland, August pp. 25-29, 2008.
12. J.-S. Lee and C. H. Park, "Adaptive Decision Fusion for Audio-Visual Speech Recognition," in *Speech Recognition, Technology and Applications*, I-Tech, Vienna, Austria, pp. 275-296, 2008.
13. J. Kennedy, and R. Eberhart, "Particle Swarm Optimization," *Proceedings of the IEEE Int. Conf. on Neural Networks*, Piscataway, NJ, pp. 1942 - 1948, 1995.
14. Kullback, S; Leibler, R.A, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. ?? (1): pp. 79 - 86, 1951.
15. A. Bhattacharyya, "On a Measure of Divergence between Two Statistical Populations Defined by Probability Distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99 - 109, 1943.
16. Printz et al., "Theory and Practice of Acoustic Confusability", *Proceedings of the ISCA ITRW ASR2000*, pp. 77-84, Paris, France, Sep. 18-20, 2000.
17. John Hershey and Peder Olsen, "Approximating the Kullback Leibler divergence between gaussian mixture models," *Proceedings of ICASSP 2007*, Honolulu, Hawaii, April 2007.

18. J.R. Hershey, P.A. Olsen, "Variational Bhattacharyya Divergence for Hidden Markov Models", *Proceedings of ICASSP 2008*, pp. 4557-4560, 2008.
19. John R. Hershey, Peder A. Olsen, and Steven J. Rennie, "Variational Kullback Leibler Divergence for Hidden Markov Models," *Proceedings of ASRU*, Kyoto, Japan, pp. 323-328, December 2007.
20. Jia-Yu Chen, Peder Olsen, and John Hershey, "Word Confusability - Measuring Hidden Markov Model Similarity," *Proceedings of Interspeech 2007*, pp. 2089-2092, August 2007.
21. J. Silva and S. Narayanan, "Average Divergence Distance as a Statistical Discrimination Measure for Hidden Markov Models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, issue 3, pp. 890-906, May 2006.
22. http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

저자 약력

•Than Trung Pham

2003.9~2007.9: Dept. of Mathematics and Computer Sciences, HCM University of Natural Sciences (BS)
2008.2~: Dept. of Electronics and Computer Eng., Chonnam National Universtiy (MS)
Research interests: machine learning, audio visual signal processing, computer vision···

•Jin Young Kim (Correspondent Author)

1986. 2: Dept. of Electronics Eng. Seoul Nat'l Univ.(BS)
1988. 2: Dept. of Electronics Eng. Seoul Nat'l Univ.(MS)
1993 .8: Dept. of Electronics Eng. Seoul Nat'l Univ.(Ph.D)
1995~: Chonnam Nat'l Univ., (professor)
Research Area: Audio-visual signal processing

•Seung Yu Na

1977.2: Dept. of Electronics Eng. Seoul Nat'l Univ.(BS)
1986: Dept. of ECE University of Iowa(Ph.D)
1987~: Chonnam Nat'l Univ. (Professor)
Research Area: Intelligent control, Signal processing