A RELIABILITY GUIDED SENSOR FUSION MODEL FOR OPTIMAL WEIGHTING IN MULTIMODAL SYSTEMS

Mustapha Makkook, Otman Basir, and Fakhreddine Karray

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

ABSTRACT

For intelligent sensory systems, it is highly desirable to develop assessment methods that can continuously evaluate the reliability of potential sensory strategies taking into consideration changes in observation conditions. This relies on measuring a set of complementary features from multiple sensors and combining these features in an "intelligent" way that maximizes information gather and minimizes the impact of noise coming from the individual sensors. In this work, we formulate a statistical assessment method for estimating the reliability of observation conditions and propose an optimal mapping into weighting measures using genetic algorithms. Our approach is particularly beneficial for multimodal systems such as audio-visual speech recognition (AVSR).

Index Terms— Multisensor systems, Reliability estimation, Speech recognition.

1. INTRODUCTION

1.1. Bayesian Fusion

Multimodal fusion or integration combines S complementary features, originating from a single or multiple modalities, in order to maximize information gather and to overcome the impact of noise in each individual stream. The simplest way to combine audio and video data is to use Bayes' rule and multiply the audio and video a posteriori probabilities. From a probabilistic perspective, this approach is valid if the audio and video data are independent. Perceptive studies have shown that in human speech perception, audio and video data are treated as class conditional independent [1]. In this case, the conditional probability of the observation vector $\mathbf{x}_{1:S} = (\mathbf{x}_1, ..., \mathbf{x}_S)$ given the class label c_i is governed by the product:

$$P(\mathbf{x}_{1:S}|c_i) = P(\mathbf{x}_1, ..., \mathbf{x}_S|c_i) = \prod_{s=1}^{S} P(\mathbf{x}_s|c_i). \quad (1)$$

Using Bayes' rule, we get the desired a posteriori probability of the class given the features:

$$P(c_i|\mathbf{x}_{1:S}) = \frac{\prod_{s=1}^{S} P(c_i|\mathbf{x}_s)}{P(c_i)} . \frac{\prod_{s=1}^{S} P(\mathbf{x}_s)}{P(\mathbf{x}_{1:S})}. \quad (2)$$

By replacing the probabilities P by estimates $\hat{P}$, we get a representation of the *Bayesian Fusion (BF)*:

$$\hat{P}_{BF}(c_i|\mathbf{x}_{1:S}) = \frac{\prod_{s=1}^{S} P(c_i|\mathbf{x}_s)}{P(c_i)} .\eta, \quad (3)$$

where the terms independent of the actual class are replaced by the normalization factor η:

$$\eta = \frac{1}{\sum_{j=1}^{M} \frac{\prod_{s=1}^{S} P(c_j|\mathbf{x}_s)}{P(c_i)}}, \quad (4)$$

where M is the number of classes. This probability can then be used in classification by making use of the *Maximum A Posteriori (MAP)* rule:

$$\hat{c} = \underset{c_i \in C}{\operatorname{argmax}} \hat{P}_{BF}(c_i|\mathbf{x}_{1:S}). \quad (5)$$

1.2. Weighted Bayesian Fusion

The standard Bayesian Fusion approach does not deal with varying reliability levels of the input streams. In order to improve classification performance, several authors have introduced stream weights $\{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_S\}$ as exponents in Equation 3, resulting in the modified score:

$$\hat{P}_{WBF}(c_i|\mathbf{x}_{1:S}) = \frac{\prod_{s=1}^{S} P(\mathbf{x}_s|c_i)^{\lambda_s}}{\sum_{j=1}^{M} \prod_{s=1}^{S} P(\mathbf{x}_s|c_j)^{\lambda_s}}. \quad (6)$$

In order to determine the weights $\{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_S\}$, we first need to define reliability measures for the individual streams. These reliability measures should reflect the quality of the observation conditions by considering statistical information conveyed in both prior and current classification results. The second step is to find an optimal mapping between these reliability indicators and the stream weights $\{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_S\}$.

This paper develops a method of modality fusion that is based on reliability. First, we propose two stream reliability indicators based on the dispersion of the a posteriori probabilities of the observation vectors. These reliability indicators are then mapped into stream weights using the genetic algorithm, in such a way that maximizes the conditional likelihood. Figure 1 shows an overall diagram of our fusion system.

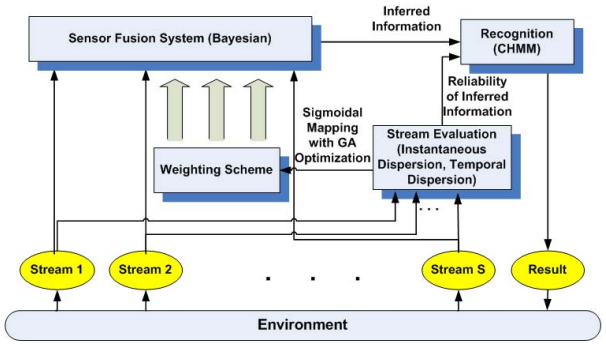

Fig. 1. Overview of the multimodal fusion system.

2. DISPERSION AS A MEASURE OF RELIABILITY

The first work to ever introduce the dispersion as a measure of reliability in audio-visual speech recognition systems was that developed by Adjoudani and Benoit [2]. Since then, this idea has been further developed by other researchers. In this work, a dispersion measure developed by Potaminaos and Neti [3] is used. This measure uses an *N*-best dispersion method that is formulated as the difference between each pair of n^{th}-best hypotheses, and it is given by:

$$L = \frac{2}{N(N-1)} \sum_{n=1}^{N} \sum_{n'=n+1}^{N} (R_n - R_{n'}), \tag{7}$$

where $N \geq 2$ and R_n is equal to the n^{th}-best hypothesis. Dispersion measures provide a good estimate of stream confidence, as a large difference in classifier outputs reflects a greater confidence. Lucey et al. [4] have theoretically proven that dispersion approximately reflects the cepstral shrinkage effect induced by additive noise.

2.1. Instantaneous Dispersion

The first reliability measure that we use is the instantaneous dispersion based on a local frame measurement:

$$\begin{aligned} L_{s,t} &= L(log(P(\mathbf{x}_{s,t}|c_{s,t}))) \\ &= \frac{2}{N(N-1)} \sum_{n=1}^{N} \sum_{n'=n+1}^{N} \log \frac{P(\mathbf{x}_{s,t}|c_{s,t,n})}{P(\mathbf{x}_{s,t}|c_{s,t,n'})}, \end{aligned} \tag{8}$$

where L(.) is the *N*-best log-likelihood dispersion function defined in Equation 7 and $P(\mathbf{x}_{s,t}|c_{s,t})$ is the observation emission probability generated by an HMM-based classifier. Here we choose $N = 4$ because both Adjoudani and Benoit [2] and Potamianos and Neti [3] have found that an *N*-best of 4 has been the most successful.

2.2. Temporal Dispersion

The instantaneous dispersion measure evaluates the stream reliability at a frame level. In highly corrupted speech, the noise will cause the dispersions to vary rapidly. Therefore, it is hard to judge whether the dispersion changes come from the varying discriminative powers of the recognizer or from the ambient noise presented in the multimodal streams. Consequently, the instantaneous dispersion is not sufficient to assess the stream reliability and there is a need for a temporal reliability measure, that takes into account the previously observed behavior of the classifier. For this purpose we propose a second confidence measure that is based on the instantaneous dispersion measure. We define this weight function $R_{s,t}$ (with backward-looking time step Δt) as:

$$R_{s,t} = \sum_{n=1}^{q} \rho(n) L_{s,t-n\Delta t} + \epsilon_t, \tag{9}$$

where q is the window length (chosen here to be equal to 5), and ϵ_t is a white noise process with zero mean and variance σ^2. The $\rho(n), n = 1, ..., q$, in the range 0 to 1, are parameters that correspond to how rapidly past performance will be discounted. This factor $\rho(n)$ is largely responsible for determining the dynamic weights $R_{s,t}$, where recent performance should be weighted highly (large $\rho(n)$) and past performance should be gradually forgotten (low $\rho(n)$). Here the problem boils down to determining the proper values of $\rho(n)$. We note from Equation 9 that this temporal dispersion is formulated as an autoregressive (AR) model. The parameters $\rho(n)$ are thus calculated using the *Yule-Walker* equations [5]:

$$r_L(m) = \sum_{n=1}^{q} \rho(n) r_L(m-n) + \sigma_\epsilon^2 \delta_m, \tag{10}$$

where $m = 0, ..., q$, yielding $q + 1$ equations, and $r_L(m)$ is the autocorrelation function of the instantaneous dispersion series. σ_ϵ is the standard deviation of the input noise process, and δ_m is the Kronecker delta function. Because the last part of the equation is non-zero only if $m = 0$, these equations are usually solved by representing them as a matrix for $m > 0$:

$$\begin{bmatrix} r_L(1) \\ r_L(2) \\ . \\ . \\ r_L(q) \end{bmatrix} = \begin{bmatrix} r_L(0) & r_L(1) & . & . & r_L(q-1) \\ r_L(1) & r_L(0) & . & . & r_L(q-2) \\ . & . & & & . \\ . & . & & & . \\ r_L(q-1) & r_L(q-2) & . & . & r_L(0) \end{bmatrix} \begin{bmatrix} \rho(1) \\ \rho(2) \\ . \\ . \\ \rho(q) \end{bmatrix}, \tag{11}$$

and we solve for all ρ. For $m = 0$ we have:

$$r_L(0) = \sum_{n=1}^{q} \rho(n) r_L(-n) + \sigma_\epsilon^2, \tag{12}$$

which allows us to solve σ_ϵ^2. Note that since the *Yule-Walker* equations are linear in the coefficients $\rho(n)$, it is a simple matter to find the coefficients $\rho(n)$ from the autocorrelation sequence $r_L(n)$.

3. STREAM WEIGHT OPTIMIZATION

The next step is to find a mapping between the reliability measures (L and R derived in Equations 8 and 9) and the stream weights $(\lambda_1, \lambda_2, \lambda_3, ..., \lambda_S)$. We use a sigmoid function for this purpose, as chosen by [6], due to the fact that it is monotonic, smooth and bounded between zero and one. First we define the reliability vector $d_t = [d_{1,t}, d_{2,t}, d_{3,t}, ..., d_{2S,t}] = [R_{1,t}, R_{2,t}, ..., R_{S,t}, L_{1,t}, L_{2,t}, ..., L_{S,t}]$.

Then, the mapping is defined as:

$$\lambda_s = \frac{1}{1 + exp(-\sum_{i=1}^{2S} w_{s,i} d_{i,t})}, \tag{13}$$

where $\mathbf{W}_s = [w_{s,1}, w_{s,2}, w_{s,3}, ..., w_{s,2S}]$ is the vector of the sigmoid parameters for stream s. Since we have S streams and $2S$ sigmoid parameters per stream, then we have $S \times 2S = 2S^2$ sigmoid parameters that we need to optimize. This is a nonlinear optimization problem with a large number of variables. Consequently, we choose to use genetic algorithms (GA) to solve this problem and determine the optimal set of weights. In our GA model, we have S streams $\{S_1, S_2, S_3, ..., S_S\}$ which require $2S^2$ parameters $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2 ..., \mathbf{W}_S\}$.

The objective function of the genetic algorithm used in this experiment is to optimize the system reliability by adjusting the weights. For this purpose we choose our objective function to be the maximum conditional likelihood (MCL) estimates of parameters $\mathbf{W}$ over the training set. Given an observation vector $\mathbf{x}_{1:S}$, we first represent the conditional likelihood of class c_i by:

$$\hat{P}_{WBF}(c_i|\mathbf{x}_{1:S}) = \frac{\prod_{s=1}^{S} P(\mathbf{x}_s|c_i)^{\lambda_s}}{\sum_{j=1}^{M} \prod_{s=1}^{S} P(\mathbf{x}_s|c_j)^{\lambda_s}}. \tag{14}$$

We then seek the parameters $\mathbf{W}$ over a time interval T in the training set as:

$$\hat{\mathbf{W}} = \underset{W}{\operatorname{argmax}} \sum_{t \in T} \log P_{WBF}(c_{i,t}|\mathbf{x}_{1:S}), \tag{15}$$

where $c_{i,t}$ is the class c_i at time t. This is subject to the constraints:

$$\sum_{s=1}^{S} \lambda_s = 1, \lambda_s \geq 0. \tag{16}$$

4. EXPERIMENTAL RESULTS

In order to evaluate the efficiency of the fusion method, classification experiments for the task of word recognition of isolated digits are conducted. The Tulips1 database is used and the audio signal is contaminated by adding 5 kinds of noise taken from the NOISEX database [7]. These different types of noise are mixed with the audio signal at 8 SNR levels ranging from -12dB to 30dB (clean speech).

Mel frequency cepstral coefficients (MFCCs) are utilized as observations for the audio stream, constructing a 26-dimensional vector. Audio-only recognition is done using HMMs. As for the visual front-end, a 12-dimensional visual feature vector is derived from the independent components of the optical flow fields. For the acoustic and visual modeling of the observations, 3-state left-right word Coupled HMMs with a single 5-continuous-Gaussian observation probability distribution per stream are used. The models are trained on clean data.

Figure 2 shows two plots. The top figure plots the instantaneous dispersions of both audio and visual streams varying over a 70-frame time window, where calculations are done at an SNR of 10dB taken from a speaker in the testing set. From this plot, it is clear that instantaneous dispersions vary dramatically over time. The bottom figure, on the other hand, plots the temporal dispersions calculated for the same speaker and same conditions. The temporal dispersion plot shows more smoothness in estimating the channel reliability than the instantaneous dispersion plot. This translates into a reliability estimate that is more robust against noise bursts.

In order to more clearly see this performance improvement, Figure 3 presents the percentage of the correctly classified words in the isolated-digit recognition task. It shows 5 curves. The "A" curve represents the audio-only speech recognition classification accuracy shown for baseline comparison. The "V" curve represents the video-only speech recognition classification accuracy. The "AV-Unweighted" curve is the baseline audiovisual setup in which we use Coupled HMMs with stream weights equal to unity for both streams. We also provide results with stream weights using the dispersion as reliability measure and the generalized gradient descent as the optimization method [6] ("AV-Dispersion" curve).

It is clear from these results that both dispersion-based fusion and fusion using the proposed approach significantly improve AVSR performance at low SNRs, with the proposed approach being somewhat superior. To further illustrate quantitatively the performance of the proposed approach to fusion, we compare fusion strategies in terms of their resulting *effective SNR gain*. We measure this gain with reference to the audio-only word classification accuracy at 10dB, by considering the SNR value where the audiovisual word classification rate equals the reference audio-only word classification rate. From Figure 3, this SNR gain is around 10dB for both the unweighted Bayesian fusion and the dispersion-based fusion. On the other hand, classification based on the proposed approach achieves a 16dB improvement, further illustrating the efficiency of this approach.

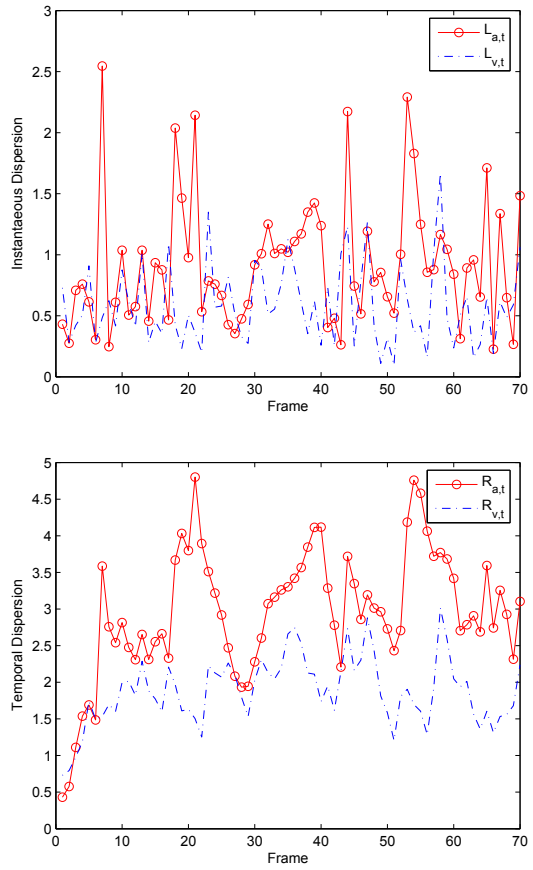

Fig. 2. Instantaneous (top) and temporal (bottom) dispersion variation at 10dB.

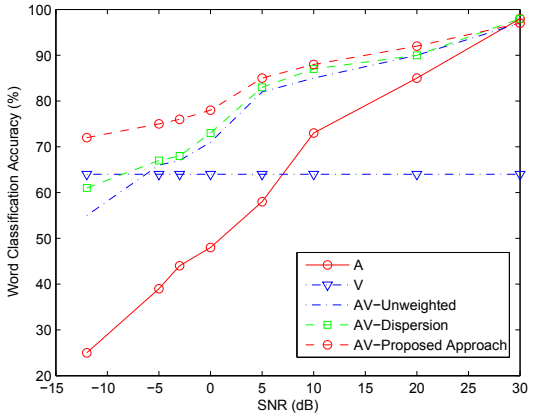

Fig. 3. Word classification accuracy.

5. CONCLUSION

In this paper, we proposed a new probabilistic reliability assessment model for multiple streams in a multimodal system. The main benefit of this assessment model is that it takes into consideration the reliability of the overall system on both a local and global level and thus is robust to sudden noise bursts. In addition, it is a model, which can be generalized for multiple information streams and multiple applications. We developed two stream reliability indicators based on the dispersion of N-best hypotheses. The reliability indicators were then mapped into stream weights using the genetic algorithm, in such a way that maximized the conditional likelihood. This optimal scheme is superior to previous approaches because it is dynamic, easy to implement, and considers an arbitrary number of streams. Experimental results did show improvements, especially at low SNR levels. Future work can extend this architecture to consider multiple streams of information on both an intramodal and intermodal level.

6. REFERENCES

[1] D.W. Massaro and D.G. Stork, "Speech recognition and sensory integration," *American Scientist*, vol. 86, no. 3, pp. 236–244, 1998.

[2] A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*, pp. 461–471, 1996.

[3] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," *Proceedings of the International Conference on Spoken Language Processing*, vol. 3, pp. 746–749, 2000.

[4] S. Lucey, Queensland University of Technology School of Electrical, and Electronic Systems Engineering, *Audio-visual speech processing*, Queensland University of Technology, Brisbane, 2002.

[5] M.H. Hayes, MH Hayes, and M.H. Hayes, *Statistical digital signal processing and modeling*, John Wiley & Sons, Inc. New York, NY, USA, 1996.

[6] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[7] A. Varga, HJM Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," *DRA Speech Research Unit, Malvern, England, Tech. Rep*, 1992.