MULTIMODAL INFORMATION FUSION USING THE ITERATIVE DECODING ALGORITHM AND ITS APPLICATION TO AUDIO-VISUAL SPEECH RECOGNITION

Shankar T. Shivappa, Bhaskar D. Rao and Mohan M. Trivedi

Department of Electrical and Computer Engineering
University of California, San Diego 9500 Gilman Drive, La Jolly, CA 92093

ABSTRACT

The fusion of information from heterogenous sensors is crucial to the effectiveness of a multimodal system. Noise affect the sensors of different modalities independently. A good fusion scheme should be able to use local estimates of the reliability of each modality to weight the decisions. This paper presents an iterative decoding based information fusion scheme motivated by the theory of turbo codes. This fusion framework is developed in the context of hidden Markov models. We present the mathematical framework of the fusion scheme. We then apply this algorithm to an audio-visual speech recognition task on the GRID audio-visual speech corpus and present the results.

Index Terms— Multimedia systems, Iterative decoding, Speech recognition, Hidden Markov models, Robustness

1. INTRODUCTION

In recent years audio visual speech recognition has emerged as a prime solution to speech recognition in drastic conditions. It is supported by the fact that speech is bimodal and by the necessity for a modality that is robust to background acoustic noise. Other human activity analysis systems have also increasingly come to rely on multimodal sensors for similar reasons. Typically, the audio modality provides information that complements the video modality. In both these cases, the information fusion scheme that combines the information from different modalities is very important.

Sensor noise affects the different modalities independently. Moreover in most real applications, the noise in each modality is non stationary. In most existing schemes, the reliability of the modality is estimated based on SNR or other measures and used in the fusion framework. This is a major disadvantage of such schemes because the quality of the modalities is in general time-varying and estimating it is non-trivial.

In this paper we present a fusion framework which is based on the theory of iterative decoding. The iterative decoding framework described here applies to HMM based recognition tasks. We demonstrate the effectiveness of the algorithm in a real world problem of audio-visual speech recognition on the GRID audio visual speech corpus [1]. This work is related to our journal submission [2], which contains more details of the iterative decoding framework along with results from synthetic problems and a simple audio-visual speech segmentation task. In this paper we extend the scope of the iterative decoding algorithm to the problem of audio-visual speech recognition.

Work described in this paper was funded by the RESCUE project at UCSD, NSF award #0331690.

2. RELATED WORK

Audio visual speech recognition (AVSR) has attracted a lot of attention from researchers in recent years. The summer workshop at JHU in 2000 [3] summarized the various approaches to construct an AVSR system, including feature extraction, feature fusion and decision fusion techniques. In [4], Nefian et. al. proposed a coupled HMM as a model for the AVSR problem. Several other schemes including discriminatory learning techniques like boosting and SVMs have been investigated to improve the efficiency of AVSR. [5] is a good overview of the main approaches to AVSR so far. As discussed in the concluding section of [5],

> ... when combining audio and visual information, a number of issues relevant to decision fusion require further study, such as the optimal level of integrating the audio and visual loglikelihoods, the optimal function for this integration, as well as the inclusion of suitable, local estimates of the reliability of each modality into this function.

In this paper we present a multimodal information fusion scheme that addresses some of these concerns, especially the inclusion of local estimates of the reliability of each modality into the fusion process. This advantage of the iterative decoding scheme is discussed in detail in [2].

3. ITERATIVE DECODING FRAMEWORK

3.1. Advantages of the iterative decoding scheme

In many applications like ASR, well-trained unimodal models might already be available. Iterative decoding utilizes such models directly. Thus, extending the already existing unimodal systems to multimodal ones is easier. Another common scheme used to integrate unimodal HMMs is the product HMM [6].The iterative decoding algorithm performs better than the joint model and the product HMM in the presence of background noise. In the case of other decision level fusion algorithms like the multistream HMMs and factorial HMMs [7] [4] and reliability weighted summation rule, one has to estimate the quality(SNR) of the individual modalities to obtain good performance. Iterative decoding does not need such apriori information. This is a very significant advantage of the iterative decoding scheme because the quality of the modalities is in general time-varying. For example, if the speaker keeps turning away from the camera, video features are very unreliable for speech recognition. The exponential weighting scheme of multistream HMMs requires real time monitoring of the quality of the modalities which in itself is a non-trivial problem.

A good fusion scheme should have lower error rates than those obtained from the unimodal models. The joint modeling framework as well as the iterative decoding framework have this property. Building joint models requires significantly greater amounts of multimodal data than training unimodal models due to the increase in dimensionality or complexity of the joint model or both. Working with unimodal models also makes it possible to use a well-learned model in one modality to segment and generate training data for the other modalities, thus overcoming the problem of lack of training data to a great extent.

3.2. Turbo codes and Iterative decoding

Turbo codes are a class of convolutional codes that perform close to the Shannon limit of channel capacity. The seminal paper by Berrou et al.[8] introduced the concept of iterative decoding. Turbo codes achieve their high performance by using two simple codes instead of a single complex code. The iterative decoding scheme is a method to combine the decisions from the two decoders at the receiver.

We draw an analogy between the fusion of decisions at the turbo decoder and the fusion of multimodal information modeled by HMMs. In turbo codes, the transmitted bits are decoded using the likelihood values from one simple code as extrinsic information to decode the same bits from the other code. The new information in these likelihood values from the second decoder are then passed back to the first decoder for improved decoding. In this paper, the iterative decoding strategy is used to decode the hidden states of a HMM based on observations from multiple modalities. Each decoder corresponds to a unimodal HMM and the likelihood of the decoded hidden states are used as extrinsic information for decoding the hidden states in the next modality. In the next section we present the mathematical details of the iterative decoding scheme.

3.3. Hidden Markov Models

Let $\omega = (A, \pi, B)$ represent the parameters of a HMM with N hidden states, that models a particular activity. Now, the decoding problem is to estimate the optimal state sequence $Q_1^T = \{q_1, q_2 \dots q_T\}$ of the HMM based on the sequence of observations, $O_1^T = \{o_1, o_2 \dots o_T\}$.

The maximum aposteriori probability state sequence is provided by the BCJR algorithm[9]. The MAP estimate for the hidden state at time t is given by $\hat{q}_t = \arg\max P(q_t, O_1^T)$. The BCJR algorithm computes this using the forward and backward recursions.
Define,

$$\begin{aligned} \lambda_t(m) &= P(q_t = m, O_1^T) \\ \alpha_t(m) &= P(q_t = m, O_1^t) \\ \beta_t(m) &= P(O_{t+1}^T | q_t = m) \\ \gamma_t(m', m) &= P(q_t = m, o_t | q_{t-1} = m'), \end{aligned}$$

$m = 1, 2 \dots N, m' = 1, 2 \dots N$

Then establish the recursions,

$$\begin{aligned} \alpha_t(m) &= \sum_{m'} \alpha_{t-1}(m') \cdot \gamma_t(m', m) \\ \beta_t(m) &= \sum_{m'} \beta_{t+1}(m') \cdot \gamma_{t+1}(m, m') \\ \lambda_t(m) &= \alpha_t(m) \cdot \beta_t(m) \\ \hat{q}_t &= \arg\max P(q_t, O_1^T) = \arg\max \lambda_t(m) \end{aligned}$$

3.4. Multimodal scenario

For the sake of clarity, let's consider a bimodal system. There are observations O_1^T from one modality and observations Θ_1^T from the other modality. The MAP solution in this case would be $\hat{q}_t = \arg\max P(q_t, O_1^T, \Theta_1^T)$. In order to apply the BCJR algorithm to this case, we can concatenate the observations (feature level fusion) and train a new HMM in the joint feature space. Instead of building a joint model, we develop an iterative decoding algorithm that allows us to approach the performance of the joint model by iteratively exchanging information between the simpler models and updating their posterior probabilities.

3.5. Iterative Decoding Algorithm

This is a direct application of the turbo decoding algorithm[8]. In this section, it is assumed that the hidden states in the two modalities have a one to one correspondence. This requirement is relaxed in the generalized solution presented in the next section.

In the first iteration of the iterative algorithm, we decode the hidden states of the HMM using the observations from the first modality, O_1^T. We obtain the aposteriori probabilities, $\lambda_t(m) = P(q_t = m, O_1^T)$.

In the second iteration, these aposteriori probabilities, $\lambda_t(m)$ are utilized as extrinsic information in decoding the hidden states from the observations of the second modality Θ_1^T. Thus the aposteriori probabilities in the second stage of decoding are given by $\Lambda_t(m) = P(q_t = m, \Theta_1^T, Z_1^T)$ where $Z_t = \lambda_t$ is the extrinsic information from the previous step. In order to evaluate Λ_t, we modify the BCJR algorithm as follows.

$$\begin{aligned} \Lambda_t(m) &= P(q_t = m, \Theta_1^T, Z_1^T) \\ \alpha_t(m) &= P(q_t = m, \Theta_1^t, Z_1^t) \\ \beta_t(m) &= P(\Theta_{t+1}^T, Z_{t+1}^T | q_t = m) \\ \gamma_t(m', m) &= P(q_t = m, \theta_t, Z_t | q_{t-1} = m') \end{aligned}$$

Then the recursions do not change, except for the computation of $\gamma_t(m', m)$. Since the extrinsic information is independent of the observations from the second modality,

$$\gamma_t(m', m) = P(q_t = m | q_{t-1} = m') \cdot P(\theta_t | q_t = m) \cdot P(Z_t | q_t = m)$$

Here $Z_t = [z_{1t} z_{2t} \dots z_{Nt}]^T$ is a vector of probability values. A histogram of each component of Z_t for $q_t = 2$ in a $N = 4$ state HMM synthetic problem is show in Figure 1. From the histogram, one can see that a simple parametric probability model for $P(Z_t | q_t = m)$ is obtained as

$$P(Z_t | q_t = m) = f(1 - z_{mt}; \beta) \cdot \prod_{i \neq m} f(z_{it}; \beta)$$

where,

$$f(x; \beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$

is an exponential distribution with rate parameter $\frac{1}{\beta}$. Other distributions like the beta distribution could also be used. The exponential distribution is chosen due to its simplicity. The rate parameter $\frac{1}{\beta}$ which is also the variance of the likelihood values is a measure of the reliability of the recognition. At each iteration, the variance of the likelihood values is estimated and used as the rate parameter $\frac{1}{\beta}$.

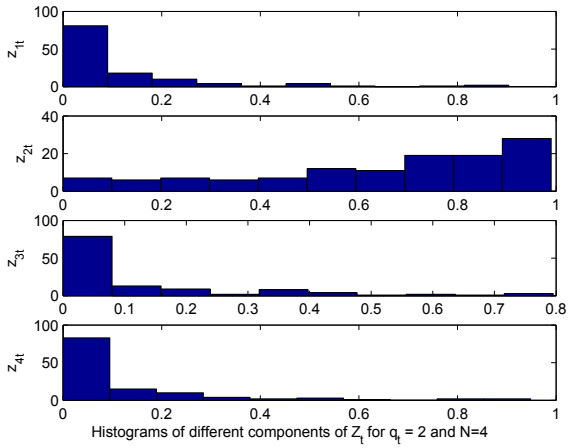

Fig. 1. A histogram of each component of Z_t for $q_t = 2$ in a $N = 4$ state HMM synthetic problem

In the third iteration, the extrinsic information to be passed back to decoder 1 is the aposteriori probabilities $\Lambda_t(m)$. But part of this information ($\lambda_t(m)$), came from decoder 1 itself. Using $Z_t = \Lambda_t$ in the third iteration destroys the independence between o_t and Z_t. From Section 3.3,

$$\begin{aligned}
\Lambda_t(m) =& \alpha_t(m) \cdot \beta_t(m) \\
\alpha_t(m) =& \sum_{m'} \alpha_{t-1}(m') \cdot \gamma_t(m', m) \\
\Lambda_t(m) =& \sum_{m'} \alpha_{t-1}(m') \cdot \gamma_t(m', m) \cdot \beta_t(m) \\
\Lambda_t(m) =& P(Z_t|q_t = m) \sum_{m'} \alpha_{t-1}(m') \\
& \cdot P(q_t = m|q_{t-1} = m') \cdot P(\theta_t|q_t = m) \cdot \beta_t(m)
\end{aligned}$$

We can thus write $\Lambda_t(m)$ as follows,

$$\Lambda_t(m) = P(Z_t|q_t = m) \cdot Y_t$$

Note that Y_t does not depend on Z_t and is hence uncorrelated with o_t . This argument follows the same principles used in turbo coding literature [8]. Hence, we normalize Y_t to sum to 1 and consider the normalized vector to be the extrinsic information passed on to the first decoder in the third iteration.

The normalized extrinsic information which is passed back to decoder 1 is given by

$$\tilde{Y}_t(m) = \frac{\Lambda_t(m)/P(Z_t|q_t = m)}{\sum_{m'} \Lambda_t(m')/P(Z_t|q_t = m')}$$

The iterations are continued till the state sequences converge in both the modalities or a fixed number of iterations are reached.

3.6. General multimodal problem

In the previous section, we assumed that the hidden states in the two modalities of a multimodal system are the same. In this section, we loosen this restriction and allow the hidden states in the individual modalities to just have a known prior co-occurrence probability. In particular, if q_t and r_t represent the hidden states in modality 1 and 2 at time t, then we know the joint probability distribution $P(q_t = m, r_t = m')$ and assume this to be stationary. This corresponds to the case where there is a loose but definite interaction between the two modalities as seen very clearly in the case of phonemes and visemes, in Audio-visual speech recognition. There is no one to one correspondence between visemes and phonemes, but the occurrence of one phoneme corresponds to the occurrence of a few specific visemes and vice-versa. We now need to compute

$$\begin{aligned}
\gamma_t(m', m) =& P(r_t = m, \theta_t, Z_t|r_{t-1} = m') \\
\gamma_t(m', m) =& P(r_t = m|r_{t-1} = m') \cdot \\
& P(\theta_t|r_t = m) \cdot P(Z_t|r_t = m) \\
\gamma_t(m', m) =& P(r_t = m|r_{t-1} = m') \cdot P(\theta_t|r_t = m) \\
& \cdot \sum_n \{P(Z_t|q_t = n) P(q_t = n|r_t = m)\}
\end{aligned}$$

This can be computed from the joint probability distribution $P(q_t = m, r_t = m')$. The rest of the iterative algorithm remains the same as before.

4. AUDIO VISUAL SPEECH RECOGNITION TASK

4.1. Database, feature extraction and modeling

We apply the iterative decoding algorithm to the AVSR problem. The GRID audio-visual speech corpus [1] is used to train the models and perform the ASR experiment. The results correspond to a speaker dependent speech recognition system. The GRID corpus is a 51 word small vocabulary speech corpus of six word long sentences. 1000 sentences are uttered by each speaker.

900 utterances are used to train the HMMs and the rest are used in the test set. Each word is modeled by a three state HMM with a Gaussian mixture model(GMM) observation density. There are 10 components in each GMM with diagonal covariance matrices. The audio feature vectors are the 13 MFCC coefficients computed on 20ms windows of audio signal with a 10ms overlap. The video rate is 25 frames per second. This corresponds to one video frame for every 4 audio frames. The video features are hence upsampled to match the audio and video frame rates. In order to extract the video features, the face of the speaker is detected and tracked using the Viola-Jones face detector[10]. The current frame is subtracted from the previous frame to estimate the motion in the mouth region of the face. The first 16 coefficients of the 2D-DCT of the mouth region motion map is used as the video feature.

4.2. Results

In the noiseless environment the audio-only speech recognizer has a state error rate of 12%. The state error rate is measured by comparing the decoded hidden state sequence with the transcriptions. The state error rate is a better estimate of the efficacy of our algorithm than the word error rate as the fusion of information takes place at the state level. The video-only speech recognizer has a state error rate of 27%. The iterative decoding algorithm converges to an error rate of 13% after the third iteration. The audio modality is then corrupted with white noise so the SNR is now reduced to 5dB. The error rate of the audio-only speech recognizer is now 40%. But the iterative

decoding algorithm converges to an error rate of 25% after the third iteration. The results are summarized in Figure 2.

In the speech recognition experiment, we do not have a joint audio-visual model to compare the performance of the iterative decoding algorithm. However, for a clearer understanding of the strengths of the iterative decoding scheme, we also present some results from simulations on synthetic data in Figure 3. In this scenario, we can build a joint model for decoding using both the modalities and we see that the iterative decoding algorithm performs better than the joint model in the presence of noise. More elaborate simulations have been presented in [2].

Note that the error rates presented here are highly dependent on the choice of the audio and video features. Using better video features would naturally lead to a better performance in the video-only speech recognizer and hence the iterative decoding framework would perform better in the presence of audio noise.

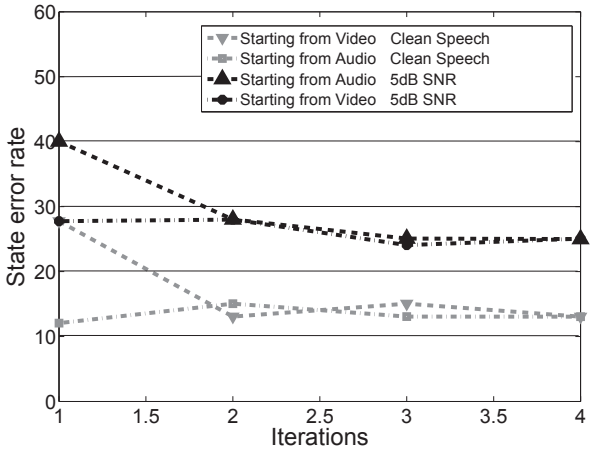

Fig. 2. State error rates for an audio-visual speech recognition task on the GRID speech corpus using the proposed scheme. After 3 iterations, the error rate of the iterative decoding algorithm converges close to the error rate of the best modality.

5. CONCLUSION AND FUTURE WORK

We have presented a multimodal information fusion scheme based on the theory of iterative decoding. The scheme has been applied to a AVSR task and the results are encouraging. The iterative decoding framework has the advantage of including local estimates of the reliability of each modality in the decoding process. This feature is especially useful in the presence of non stationary background noise. The scheme is also directly applicable to other multimodal systems which use HMMs, as is often the case with human activity analysis systems. In future, we plan to extend the iterative decoding framework to a broader range of multimodal systems.

6. REFERENCES

[1] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, pp. 2421–2424, Nov 2006.

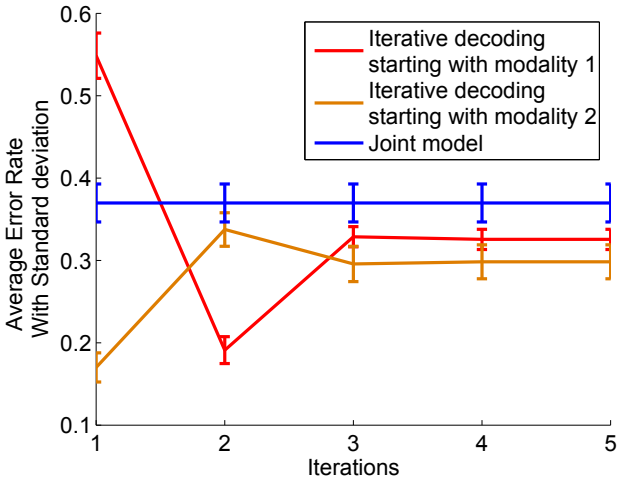

Fig. 3. Error rate at different iterations in the case of noisy modalities for a synthetic problem. Note that the iterative algorithm performs better than the joint model when one of the modalities is noisier (in this case, modality 1) than the other.

[2] S. Shivappa, B. Rao, and M. Trivedi, "An iterative decoding algorithm for fusion of multimodal information," *EURASIP Journal on Advances in Signal Processing - special issue on Human-Activity Analysis in Multimedia Data*, Jan 2008.

[3] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins summer 2000 workshop," in *Proceedings of IEEE Workshop Multimedia Signal Processing*, Cannes, 2001.

[4] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1274–1288, 2002.

[5] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, *Issues in Visual and Audio-Visual Speech Processing*, chapter Audio-Visual Automatic Speech Recognition: An Overview, MIT Press, 2004.

[6] J. Huang, Z. Liu, and Y. Wang, "Integration of multimodal features for video scene classification based on hmm," in *Proceedings of IEEE Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, 1999.

[7] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, Sept. 2000.

[8] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near shannon limit error-correcting coding and decoding:turbo-codes," in *Proceedings of the IEEE International Conference on Communications*, Geneva, Switzerland, May 1993.

[9] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Transactions on Information Theory*, vol. IT-20(2), pp. 284–287, Mar. 1974.

[10] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, 2002.