DBN BASED MULTI-STREAM MODELS FOR AUDIO-VISUAL SPEECH RECOGNITION

John N. Gowdy, Amarnag Subramanya

Clemson University
Clemson, SC - 29634.
{jgowdy,asubram}@clemson.edu.

Chris Bartels, Jeff Bilmes *

University of Washington
Seattle, WA - 98195.
{bartels,bilmes}@ee.washington.edu.

ABSTRACT

In this paper, we propose a model based on Dynamic Bayesian Networks (DBNs) to integrate information from *multiple* audio and visual streams. We also compare the DBN based system (implemented using the Graphical Model Toolkit (GMTK)) with a classical HMM (implemented in the Hidden Markov Model Toolkit (HTK)) for both the single and two stream integration problems. We also propose a new model (mixed integration) to integrate information from three or more streams derived from different modalities and compare the new model's performance with that of a synchronous integration scheme. A new technique to estimate stream confidence measures for the integration of three or more streams is also developed and implemented. Results from our implementation using the Clemson University Audio Visual Experiments (CUAVE) database indicate an absolute improvement of about 4% in word accuracy in the -4 to 10db average case when making use of two audio and one video streams for the mixed integration models over the sychronous models.

1. INTRODUCTION

In recent years, the task of noise robust automatic speech recognition has become an active topic with a number of techniques being proposed to improve word accuracies in difficult environments. The use of multi-stream models is one such technique [1]. The streams may be multi-modal (audio and visual), or simply different sets of features (MFCC, RASTA etc.) extracted from the same speech data. In particular, streams that have complimentary information have been used to improve recognition accuracies at low SNR's [2, 3]. However, one of the problems associated with using more than one stream is the need for efficient models to combine them. The goal of the fusion process should be to avoid catastrophic fusion (the combined stream performance is worse than either of the streams used independently).

The approaches that have been used for integration of two or more streams may be classified into three categories: feature fusion (or early integration), decision fusion (or late integration) and model fusion. Early integration makes the assumption that the streams are synchronous, whereas in case of late integration the streams are allowed to develop independently over time [4]. In case of model fusion, various heuristic-based combination strategies are used to form a unified HMM model from separately trained HMMs [3]. At this time, model fusion seems to be the best technique to integrate information from two streams.

*The authors Bartels and Bilmes were supported for this work by NSF grant IIS-0093430 and by an Intel Corporation grant

Another issue related to integration of information from two or more sources is the use of stream exponents. In cases, where different streams have similar performance (word accuracies) under similar conditions (SNR), near optimum output may be obtained without the use of stream exponents. However, when we are combining two streams that have dissimilar performance, the use of stream exponents is extremely important. For example, when we are trying to integrate a visual stream and an audio stream, the performance of the visual stream is independent of audio SNR, and the performance of audio stream varies directly with SNR, hence failure to use SNR dependent stream exponents under such circumstances can lead to reduced word accuracies.

In this work, we propose the use of DBN based models to combine audio and visual streams. We also propose a mixed type of DBN that can handle the integration of two or more streams and suggest a technique to estimate stream exponents for multi-modal multi-stream models with more than two streams. In section 2, we describe multi-stream DBN models, section 3 describes the experimental setup and the results are discussed in section 4. The implications of the results and future work are discussed in section 5.

2. MULTI-STREAM DBN MODELS

A Bayesian Network is a statistical model that can be used to represent collections of random variables and their dependencies. A DBN is used to model random variables as they evolve over time (e.g. as in speech). In this work, we use the graphical model structure for continuous speech given in [5] as our baseline model. The goal of this work is to extend the baseline model for audio-visual speech recognition in cases where more than two streams from independent sources are involved.

2.1. Synchronous and Asynchronous Models

A synchronous model assumes that all the streams are derived from synchronous sources of information. This assumption is valid when modeling information from the same modality of speech that has been processed in different ways (e.g. MFCC, RASTA, PLP, etc.). When we are combining information from two different modalities (such as audio and video), this assumption is not necessarily always true. It has been demonstrated that there can exist a degree of asynchrony between the audio and visual modalities [6], and representing this asynchrony in some cases can be beneficial. Further studies of this phenomena are given in [7].

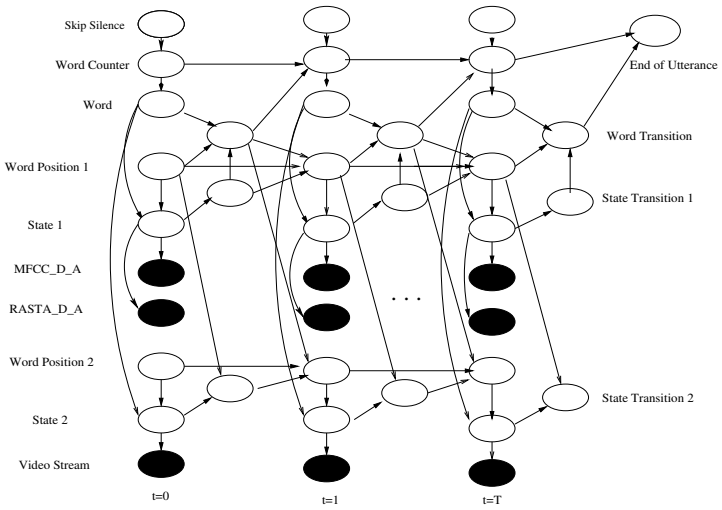

Fig. 1. Mixed Type Multi-Stream Model

2.2. Mixed Models

The structure of our new mixed model is given in Figure 1. It can be seen that it models streams processed from the voice input using a synchronous structure whereas the video stream is modeled asynchronously with the 'composite' audio stream. In cases where we make use of a single audio stream and a single video stream the mixed model reduces to an asynchronous multi-stream model. Thus, the mixed model may be considered as a combination of a synchronous and asynchronous multi-stream DBN.

In the synchronous part of the model, all the observation variables share one state variable, so that all audio streams are synchronized at the state level. However, the video stream and composite audio stream each depend on different state variables which introduces the asynchrony between them. In order to model the synchrony between the audio and video streams, they are made to share one word variable thereby requiring that the two streams be synchronized at the beginning of each word. It should be noted that the variable `State_Transition_2` is modeled as a child of both `State_2` and `Word_Position_1`. This is a random dependency, not a deterministic one. The intent of modeling this relationship is to limit the asynchrony between the two streams which encourages, but does not require, the second stream to also be synchronized at the end of each word. A complete description of the variables used in the model their and CPD's (conditional probability distributions) may be obtained in [7].

3. EXPERIMENTAL SETUP

3.1. Database Description

We have used the Clemson University Audio-Visual Experiments (CUAVE) database for all work in this paper. It consists of 36 speakers (19 male and 17 female) speaking digits in a connected fashion. The video stream consists of frontal images of the speakers with no rules regarding the position of the speaker within a given frame. A detailed description of the database is given in [8].

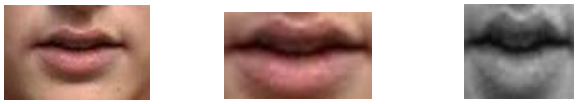

Fig. 2. Figure showing the three stages during extraction of the speaker's mouth region.

3.2. Audio Feature Extraction

We have made use of the HTK feature extraction program to extract MFCC features. The speech input was processed using a 30ms Hamming Window, with the frame period set to 10ms. For each frame 13 MFCC features were extracted; delta and acceleration coefficients were appended to the MFCC features resulting in a 39 dimensional MFCC_D_A feature vector. We have used a toolkit from ICSI/UC Berkeley to generate the RASTA features [9]. The RASTA features were also processed in HTK format to generate RASTA_D_A features.

3.3. Visual Feature Extraction

The extraction of visual features starts with the detection of the speaker's eyes. The eye detection algorithm is described in [10]. Once the position of the speaker's eyes are obtained, we make use of the distance between the speaker's eyes to estimate the approximate position of the speaker's mouth region (shown in first image in Figure 2). Next, in order to obtain a more accurate fix on the speaker's mouth region we make use of Linear Discriminant Analysis (LDA) [11] to classify the pixels in the mouth region into lips and non-lips. The HSI (Hue-Saturation-Intensity) color space is used as input to the LDA stage. The optimal linear discriminant is computed off-line using a set of manually segmented images of the speaker's mouth region from the CUAVE database. The LDA stage results in a better estimate of the speaker's mouth region. The results of the LDA stage are shown in the second image in Figure 2

In order to make the lip features rotation invariant, we make use of principal component analysis (PCA) to estimate the angle of rotation of the speaker's mouth region. The first two eigenvectors (the two vectors with the largest eigenvalues) obtained from a PCA computation of the mouth region are used to estimate the rotation angle and then the mouth region is corrected using an affine transformation. We then down-sample the mouth region to a 16×16 gray-scale intensity image, and then a 2D-discrete cosine transform is applied. The last image in Figure 2 is the speaker's mouth region before the DCT is applied (it has been magnified for display purposes). The upper 30 coefficients of the resulting computation are retained (the DC value is also discarded.). Delta coefficients are appended to the static video feature vector resulting in a 60 dimensional video feature vector.

The video features are extracted at 25Hz. Since, the audio features were processed at 100Hz, the video features were interpolated to make them occur at the same frame rate as the audio features.

3.4. Setup

The CUAVE database was divided into a testing set and a training set. The testing set consisted of 12 speakers and the training set consisted of 26 speakers, the speakers in each of the sets were chosen randomly (two speakers are common to the testing and train-

SNR(db)	-4	4	6	10	12	Clean
λ_a	0.3	0.55	0.70	0.75	0.85	0.95

Table 1. Audio Stream Exponents

ing sets). Our first goal was to compare the word accuracies for GMTK and HTK under similar conditions when making use only of a single stream (either audio or video). The audio stream made use of MFCC_D_A features and the video stream features were extracted as described in section 3.3. The models were trained on clean speech and then tested at various SNRs ranging from -4db to 12db (mis-matched condition).

Our second setup was to compare the word accuracies for GMTK and HTK when making use of two streams i.e. an audio stream and a video stream. In this case, since we are making use of two streams with inherently different word accuracies, stream exponents can be used to achieve the best balance between the two streams, leading overall to the best accuracy improvement. The multi-stream system implemented in HTK is described in [3]. We have made use of the same stream exponents for both the HTK and GMTK setups. The audio stream exponents for each of the SNR's are as shown in Table 1. The video stream exponents may be computed from this table using $\lambda_v = 1 - \lambda_a$. The stream exponents were estimated by making use of a reduced set. The reduced training set and testing sets for the stream weight estimation were obtained from the actual training (10 speakers) and testing sets (3 speakers). For a given SNR ($-4db$ to $12db$), λ_a was varied from 0 to 1 in steps of 0.05, and the value of the stream exponent that maximized the word accuracy was chosen. The parameters for the asynchronous two-stream GMTK system were first bootstrapped from the single stream models and then were jointly trained using a few additional EM iterations.

The third setup is to model the integration of more than two streams using DBN's. All parameters for the asynchronous and mixed models were bootstrapped from single stream models. One of the main issues associated with implementing such a multi-stream model is the estimation of stream exponents for each of the streams. We adjust the stream weights such that they obey the following constraint, $\lambda_{MFCC} + \lambda_{RASTA} + \lambda_v = 1$, where λ_{MFCC}, λ_{RASTA} and λ_v are the stream exponents for the MFCC, RASTA, and video streams respectively. Until now, no algorithm has been proposed to estimate the above parameters, however [12] is related to this problem. We could use the same technique that was used to estimate stream exponents in the case of two streams; however, in this case we would have two free variables even after one of them is fixed. Hence that approach would be tedious and inefficient. In this paper, as a simple first attempt, we have made use of an ad-hoc approach to estimate the stream exponents for the three stream case. They are derived from the two stream case. For any given SNR, we assume that the λ_v in the three stream case is essentially the same as the λ_v in the two stream case and $\lambda_{MFCC} = \frac{\lambda_a}{2}$, where λ_a is the audio stream exponent for the two stream case. Although, these might not be the optimum stream exponent values, they serve as good starting points for investigating the feasibility of mixed models for information fusion. In section 5 we suggest an approach that could be used to estimate the stream exponents in the three stream case. In all models, each stream is modeled by 16 states per word (whole word models) and 4-mixtures per state model. All the DBN based models were implemented using GMTK [13]. The GMTK system was implemented on a Beowulf Parallel Computing cluster with 16 nodes.

4. RESULTS

The results for the single stream case are given in Table 2. It can be seen that on the testing set, at $SNR = -4db$, there is an improvement of 6.46% in the word accuracy for the GMTK system over the HTK system. Overall, ($SNR = -4db$ to $12db$), the GMTK system shows an improvement of about 3.3%. Another interesting aspect of these results is that the improvement in word accuracies is more pronounced in cases of low SNRs. The GMTK video only recognizer performs slightly worse than the HTK video only recognizer. A reason for this could be the use of the MFCC training schedule (mixture-coefficient vanishing ratios and mixture co-efficient split ratios) for training the video stream.

The two stream integration results are as shown in Table 3. Here we compare the performance of similar HTK and GMTK systems when fusing the audio and video streams. It can be seen that the use of DBN's to model two streams outperforms the HTK system, by about 5% at an SNR of $-4db$ on the testing set for the synchronous case. Overall, (averaged over $-4db$ to $12db$) the GMTK system outperforms the HTK system by about 2%. Also shown in Table 3 are the two stream asynchronous model results for the GMTK system. It can be seen that the asynchronous system gives a better word accuracy when compared to the synchronous approach. This suggests that the best means of modelling the audio and visual streams is to make use of asynchronous modelling.

Shown in table 4 are the results for the three stream case. In this setup all the systems have been implemented in GMTK alone. The entirely synchronous models perform better than the 3-stream asynchronous models. This is similar to the results obtained in [7], the reason being that the two stream MFCC and RASTA are inherently synchronous and we are forcing them to be modeled by an asynchronous structure. However, the mixed integration scheme outperforms both the synchronous and the asynchronous types of integration. It can be seen that there is an improvement of about 4% in word accuracy for the mixed models over the synchronous multistream models at an SNR of $-4db$ on the testing set.

5. CONCLUSIONS AND DISCUSSIONS

In this paper we have described techniques to combine two or more streams of information for speech recognition. Specifically, using GMTK we implemented and tested several multi-stream DBN models for speech that incorporated multiple acoustic and visual features. Results show that the use of DBN's leads to significant improvement in word accuracies. Hence the use of DBN's is a simple and effective way of combining information from two or more modalities. We have also shown that when combining information from a number of sources, we need to account for the inherent synchrony and asynchrony between the modalities. This point is illustrated by comparing the results in the two stream and three stream cases for the asynchronous models, in the two stream case the asynchronous models perform better than synchronous models whereas in the case of three stream models the converse is true. The mixed models best account for the synchrony and asynchrony between the audio and visual streams and hence perform the best.

However, an issue related to the multi-stream models (more than two streams), is the estimation of stream exponents. In order to estimate the stream exponent we could train a two stream HMM on a reduced training set with one of the streams being the

Setup	-4db	4db	6db	10db	12db	Clean	-4 to 12db
Audio Only MFCC_D_A **(HTK)**	31.00	65.57	75.67	84.82	89.33	98.00	69.28
Audio Only MFCC_D_A **(GMTK)**	37.46	71.33	79.34	84.33	90.41	97.92	72.57
Video Only **(HTK)**	53.33	53.33	53.33	53.33	53.33	53.33	53.33
Video Only **(GMTK)**	51.40	51.40	51.40	51.40	51.40	51.40	51.40

Table 2. Single Stream Results: Word Accuracies (in %) for the HTK and GMTK systems

Setup	-4db	4db	6db	10db	12db	Clean	-4 to 12db
Two Stream Synchronous MFCC_D_A and Video **(HTK)**	63.13	76.67	82.33	89.11	92.33	98.33	80.71
Two Stream Synchronous MFCC_D_A and Video **(GMTK)**	68.33	79.67	82.33	88.11	94.72	98.64	82.63
Two Stream Asynchronous MFCC_D_A and Video (GMTK)	**72.12**	**82.67**	**84.46**	**90.01**	**96.12**	**98.92**	**84.96**

Table 3. Two Stream Results: Word Accuracies (in %)

Setup	-4db	6db	10db	Clean	-4 to 10db
Three Stream Synchronous MFCC_D_A, RASTA and Video	66.12	84.14	90.11	98.10	80.12
Three Stream Asynchronous MFCC_D_A, RASTA and Video	66.10	79.12	87.40	97.92	77.54
Three Stream Mixed MFCC_D_A, RASTA and Video	**70.14**	**88.12**	**94.11**	**99.31**	**84.12**

Table 4. Three Stream Results: Word Accuracies (in %)

composite MFCC and RASTA features and the other stream being the video stream. This setup could be used to estimate λ_v and $\lambda_{MFCC} + \lambda_{RASTA}$. We could then construct another two stream HMM, once again using a reduced training set, with one stream being MFCC and the other being RASTA. This setup could be used to estimate λ_{MFCC} and λ_{RASTA}, with the constraint that their sum must be equal to the value estimated above. Hence, we are essentially decomposing the stream estimation procedure to handle two streams at any given time, as the stream exponents for the two stream case may be easily estimated.

6. REFERENCES

[1] A. Janin, D. Ellis, and N. Morgan, "Multi-stream speech recognition:ready for prime time," in *Proc. of Eurospeech*, Budapest, 1999.

[2] C. Neti, G. Potamianos, J. Luettin, I. Matthews, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio visual speech recognition," in *Final Report: JHU 2001 Summer Workshop*, 2000.

[3] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition," in *Proc. of IEEE International Conf.on Acoustics, Speech and Signal Processing*, Orlando, Florida, 2002.

[4] J.G. Fiscus, "A post-processing system to yeild reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, California, 1997.

[5] J. Bilmes, G. Zweig, and et. al., "Discriminatively structured dynamic graphical models for speech recognition," in *Final Report: JHU 2001 Summer Workshop*, 2001.

[6] A.V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and Kevin Murphy, "A coupled HMM for audio-visual speech recognition," in *Proc. of IEEE International Conf.on Acoustics, Speech and Signal Processing*, Orlando, Florida.

[7] Y. Zhang, Q. Diao, S. Huang, W. Hu, C. Bartels, and J. Bilmes, "DBN based multi-stream models for speech," in *Proc. of IEEE International Conf.on Acoustics, Speech and Signal Processing*, Hong Kong, 2003.

[8] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. of IEEE International Conf.on Acoustics, Speech and Signal Processing*, Orlando, Florida, 2002.

[9] H. Morgan, N. Bayya, A. Kohn, and P. Hermansky, "RASTA-PLP speech analysis," in *ICSI Technical Report TR-91-069*, Berkeley, California, 1991.

[10] S. Amarnag, R. Kumaran, and J. N. Gowdy, "Real time eye-tracking for human computer interfaces," in *Proc.of IEEE International Conf. on Multimedia and Expo.*, Baltimore, Maryland, 2003.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley Sons Inc., New York, NY, 2000.

[12] Dimitri Vergyri, "Use of word level side information to improve speech recognition," in *Proc. of IEEE International Conf. on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000.

[13] Jeff Bilmes and Geoffrey Zweig, "The graphical models toolkit: An open source software system for speech and time-series processing," in *Proc. of IEEE International Conf.on Acoustics, Speech and Signal Processing*, Orlando, Florida, 2002.