USING MULTIPLE VISUAL TANDEM STREAMS IN AUDIO-VISUAL SPEECH RECOGNITION

Ibrahim Saygin Topkaya and Hakan Erdogan

Vision and Pattern Analysis Laboratory
Sabanci University - Faculty of Engineering and Natural Sciences
Istanbul, Turkey
{isaygint,haerdogan}@sabanciuniv.edu

ABSTRACT

The method which is called the "tandem approach" in speech recognition has been shown to increase performance by using classifier posterior probabilities as observations in a hidden Markov model. We study the effect of using visual tandem features in audio-visual speech recognition using a novel setup which uses multiple classifiers to obtain multiple visual tandem features. We adopt the approach of multi-stream hidden Markov models where visual tandem features from two different classifiers are considered as additional streams in the model. It is shown in our experiments that using multiple visual tandem features improve the recognition accuracy in various noise conditions. In addition, in order to handle asynchrony between audio and visual observations, we employ coupled hidden Markov models and obtain improved performance as compared to the synchronous model.

Index Terms— Audio-Visual Speech Recognition, Hidden Markov Models, Tandem Approach, Support Vector Machines, Neural Networks, Coupled Hidden Markov Models

1. INTRODUCTION

Conventional speech recognition with hidden Markov models (HMM) [1] processes audio data using hidden state machines with Markovian transitions and Gaussian mixture emissions. Since audio channel noise is an important factor that affects recognition accuracy negatively, audio data may be processed so that it is less sensitive to noise or supported with visual data to increase accuracy.

In order to increase audio feature robustness, HMM's generative modeling of the observation data can be supported with discriminative classifiers where outputs of the classifiers are used as observation features, resulting in a tandem HMM system [2] .

Also, supporting the audio information with visual information is another popular technique. Usually the visual features are handled in a separate stream in a multi-stream HMM (MSHMM) [3]. However, since regular multi-stream HMM handles both channels synchronously and there may be asynchrony between audio and video channels, some extensions like coupled hidden Markov models (CHMM) [4], product hidden Markov models (PHMM) [3] and multi-stream asynchrony dynamic Bayesian networks [5] have been proposed to take this asynchrony into consideration. Also as investigated in [6], expressions *product HMM* and *coupled HMM* are sometimes referred interchangeably throughout the literature.

In this work we propose architectures of multi-stream and coupled HMMs, which uses both direct audio-visual observation features and visual tandem features extracted from support vector machine (SVM) and neural network (NN) classifiers. Neural networks, particularly multilayer perceptrons were successfully used in tandem speech recognition studies before [2]. SVMs were also employed with success in speech recognition as well [7]. These two different classifiers have strong and complementary properties which makes them good candidates for extracting separate streams of posterior probabilities. According to the best of our knowledge, it is a novel idea to use multiple classifier posteriors as separate streams in speech recognition. Conventionally, classifier combination may be performed at the frame level by decision fusion of individual classifiers' posterior probabilities, however we show that one can have much higher improvement by using model-level fusion through the use of multiple streams. We have only used visual tandem streams since for our problem of interest we have not had much improvement with additional audio tandem streams. However, audio tandem streams may also be incorporated to the system to improve accuracies in general. In addition, we implement the CHMM which allows asynchrony of audio and visual streams with a modified multi-stream tied HMM model [8] in this work. This implementation enables efficient initialization and training procedures for the CHMM and yields much improved results in accuracy.

We organize the rest of the paper as follows. In the next section, we discuss the tandem approach in speech recognition. In section 3, we describe the multi-stream and coupled HMMs. We present how a coupled HMM can be represented as a stream-tied MSHMM in section 4. Section 5 gives the details about the experimental framework and the architecture of the system. Results of the experiments are presented and discussed in section 6. Finally, we present our conclusions in section 7.

2. TANDEM FEATURES FOR SPEECH RECOGNITION

Using posterior probabilities of a classifier as a feature vector is a well known technique in speech recognition [2]. The idea in this "tandem approach" is adding a classifier layer after feature extraction. The class definition for the classifier can be chosen parallel to the HMM, such that each class can be one of words, sub-words, phones, monophone states or context-dependent phone states. For example, consider a monophone HMM model for digit recognition with ten words (one for each digit), around twenty phones (depending on the language) and a total of sixty monophone states. The

This research is supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under the scientific and technological research support program (code 1001), project number 107E015 entitled "Novel Approaches in Audio Visual Speech Recognition".

tandem classifier may be trained to discriminate one of these units of the model.

The outputs of the classifier are then considered in the HMM model as observation vectors. Usually the values are directly used, so for a tandem classifier trained for a number of C classes, HMM observations are vectors of length C.

Although originally proposed for audio-only speech recognition, the idea can be used for video based features as well by applying the same process to extracted video features. In this work, we employ the tandem approach for video data, using multiple classifier outputs in addition to regular observation features. We model all streams of data using a MSHMM where each stream comes from different sources; one for audio features, one for video features and two sets of features extracted from tandem classifiers.

3. MULTI-STREAM AND COUPLED HMM

In a multi-stream observation sequence $(\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$, we assume that each observation is $\boldsymbol{o}_t^T = [\boldsymbol{o}_t^{1^T}, \ldots, \boldsymbol{o}_t^{S^T}]$, where S is the number of modeled streams. The emission probability for a state q_t is calculated by:

$$p(\boldsymbol{o}_t|q_t) = \prod_{i=1}^{S} p(\boldsymbol{o}_t^i|q_t)^{\lambda_i}. \quad (1)$$

Here λ_i are the stream weights. In this model, the probability of transition from state j to i is given as:

$$p(q_t = i|q_{t-1} = j). \quad (2)$$

However, one problem in this assumption is that, real world data are not always in perfect synchrony. In an audio-visual system, this asynchrony can be due to the nature of the speech generation process or because of small delays in audio-visual data acquisition and processing. An example is in generation of plosives like the phone "p" where the lip position changes before the hearing of the plosive sound. So, one should consider modeling the difference of "generation timing" across the modalities.

One extension to MSHMM is coupled HMM in which independent transitions of streams are allowed. In this model, the probability of transition for one stream depends on previous states of both streams:

$$p(q_t^1 = i|q_{t-1}^1 = i', q_{t-1}^2 = j'), \quad (3)$$

where q_t^1 and q_t^2 represent individual states corresponding to the first and the second streams for time t.

The difference between MSHMM and CHMM can be seen visually by comparing the graphical models of the MSHMM and CHMM in Figure 1 (a), where squares represent hidden states and circles represent observations.

4. MODELING CHMM AS A STREAM-TIED MSHMM

Although being different models, a CHMM can be equivalently modeled as a stream-tied MSHMM as proposed in [8] by adding hybrid states to the MSHMM model, where these new states have stream-level tying of their emission probabilities.

For example, considering the case of a two stream MSHMM where audio stream has moved to another state where $q_t^1 \neq q_{t-1}^1$ but video stream has stayed at the earlier state that is $q_t^2 = q_{t-1}^2$ which is easily modeled in CHMM as in equation (3) is not directly handled by an MSHMM. However this behavior can be incorporated into an

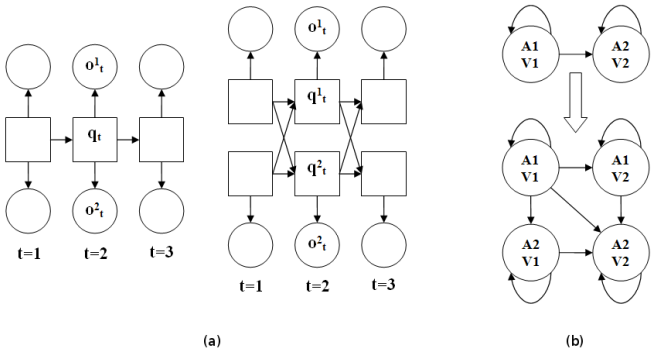

Fig. 1. (a) Graphical model of MSHMM (left) and CHMM (right) (b) Converting a left-to-right two state MSHMM to a four state stream-tied MSHMM.

MSHMM by considering joint or hybrid states $q_t = (i, j)$ in the model, where observation probabilities are derived by cloning audio stream of state i and video stream of state j. The emission likelihood of the new state is given as:

$$p(\boldsymbol{o}_t|q_t = (i,j)) = p(\boldsymbol{o}_t^1|q_t^1 = i)^{\lambda_1} p(\boldsymbol{o}_t^2|q_t^2 = j)^{\lambda_2}, \quad (4)$$

where q_t^i represent stream states, q_t is the joint state and λ_1 and λ_2 are the stream weights. The transition probabilities between hybrid MSHMM states are related to the CHMM model by the following formula:

$$p(q_t = (i,j)|q_{t-1}) = p(q_t^1 = i|q_{t-1}) p(q_t^2 = j|q_{t-1}). \quad (5)$$

This model can be easily generalized for more than two asynchronous streams as well.

This simple example demonstrates how asynchronous transitions of streams can be handled by deriving a new model from an existing one. To generalize the idea, a MSHMM with Q states and S streams/channels can be converted to a model behaving like a CHMM having a total of Q^S states. Typically, we allow asynchrony inside an HMM model only, so in a large vocabulary the increase in the number of states is proportional to the average number of states in a model (such as the model of a single phone) and not to the total number of states. Figure 1 (b) also visualizes the process using state transition diagrams.

4.1. Training the Stream-tied MSHMM

We first train a regular MSHMM model and then form hybrid states from them. The emission probabilities of the hybrid states are initialized using the original MSHMM states' emission probabilities. Since the original MSHMM describes transitions only between the original Q states, the transitions involving the new states cannot be inferred directly from them. One can initially give equal probabilities to transitions:

$$p(q_t = (i,j)|q_{t-1} = (i',j')) = 1/4, \quad (6)$$

where $0 \leq i - i' \leq 1$ and $0 \leq j - j' \leq 1$ since we deal with a left-to-right MSHMM with no skips.

We have experimented with two different training strategies in our work. The first training strategy (method 1) only updates the transition probabilities in the stream-tied MSHMM. The second

training strategy (method 2) updates both the transition and emission probabilities (in a tied fashion) during training. Thus, the second training approach enables asynchronous learning of the emission distribution parameters as well.

In our work we model a MSHMM system, where one stream is audio data, and other three streams are visual data. Since asynchrony is a problem only between data acquired from different sources, we take it into consideration between one audio stream and the group of three video streams.

5. ARCHITECTURE OF THE TANDEM-CHMM SYSTEM

We examine the proposed model with the M2VTS video database [9], which consists of five videos of 37 different people recorded in different times and arranged in five tapes. We have used first four tapes as training data and the fifth tape as testing data, since first four tapes are recorded under similar conditions and for each subject fifth tape has some visual differences (e.g. glasses, hat) that add extra difficulty to the test set. On the videos, the speakers say ten French digits so we have ten words (digits) and 19 phonemes.

For audio data Mel frequency cepstral coefficients (MFCC) [10] with 13 static plus Δ and $\Delta\Delta$ features are used for each window. For visual data, lip region from each frame is extracted as region of interest (ROI) using the method proposed in [11]. After the ROI is extracted for the whole video, Principle Component Analysis (PCA) is applied on frames, extracting top 30 principle components for each frame. Combined with first derivatives over time dimension, a frame is represented with a vector of 60 dimensions.

The class of each frame is determined using alignment done from an HMM trained on only clean audio data of the videos since clean audio is the most reliable data and we use it as a baseline for our experiments. Input to tandem classifiers are obtained by splicing 9 consecuive frames resulting in 540 features. We take phones as classes; since using words result in a small number of complex classes and using phone states would result in too many classes. So tandem classifiers discriminate 19 classes, and generate feature vectors of length 19 where each dimension corresponds to the output of the classifier for each class. We use stacked generalization with four-fold cross-validation to train the tandem classifiers and extract posterior probabilities [12].

We train two different tandem classifiers; one using NN and one using SVM algorithms. Both classifiers output normalized continuous values and can be directly used like an observation vector in a continuous HMM and because of this, both classifiers have been preferred in speech recognition systems [2, 7]. We have selected parameters that give the best frame-level cross-validated classification accuracy; NN classifier having one hidden layer consisting of 100 neurons, and SVM classifier using Radial Basis (RBF) kernel with parameter values $C = 0.5$ and $\gamma = 2^{-5}$.

After we get features of length 19 from each classifier, we concatenate them with their first and second derivatives (i.e. Δ and $\Delta\Delta$) resulting in 57 features. To handle these features in the MSHMM better, we use maximum likelihood linear transform (MLLT) [13] which tries to linearly transform data to a space where the class-dependent likelihood of the data under a diagonal-covariance modeling assumption is maximized.

To simulate noisy recording conditions, we have added noise in different SNR levels to the audio signal, and trained models using only clean audio. The noise is the Volvo 340 car noise obtained from the NOISEX database [14] and the SNR levels are determined using the "audio voltmeter" program from the G.191 ITU-T STL software suite [15].

Our MSHMM consists of four streams; (1) audio, (2) visual, (3) visual tandem using SVM classifier and (4) visual tandem using NN classifier. The contribution of each stream to the decoding process differs on each experiment; we examine different stream weights between 0 and 1, in steps of 0.25. We use stream-tied MSHMM model equivalent to a CHMM as proposed in section 4. First we model the phones with three states, and train an audio-only HMM. To obtain an audio-visual MSHMM, we concatenate visual data to audio data and train the multi-stream model using single-pass retraining from the audio-only HMM with only one iteration. Next, to create an initial model for the CHMM, we couple the states by adding the hybrid states during which we take audio stream alone and couple it with the remaining streams (since all of them are derived from the same visual channel) thus resulting in nine states for each phone. Then using the state coupled regular MSHMM model as the initial model we apply two different training methods as discussed in section 4.1 to train stream-tied MSHMMs equivalent to the CHMM. Stream weights are given as one during stream-tied training.

6. RESULTS AND DISCUSSION

For each SNR level we get the results by trying out different combinations of stream weights as mentioned in section 5. At each SNR level we present four different recognition accuracy rates shown in Table 1 for regular synchronous MSHMM and CHMM obtained with two different training methods proposed in section 4.1. For the columns labeled *"Audio"* and *"Video"*, only audio or video stream is active by giving zero weights to the unused streams and one to the used stream. For the *"Audio Visual"* column, the audio and video weight combination that gives the best result is utilized, with zero weights for tandem streams. For the *"AudioVisual and Tandem"* column, the weight combination among all four streams that gives the best result is used for each SNR value.

The results clearly show that, as the SNR decreases (i.e. noise increases) the weight combinations that emphasize visual data tend to give better results (can be seen by comparing audio only results with video only or audiovisual results). This is a well known result in audio-visual speech recognition, since due to audio noise, contribution of audio channel to the accuracy decreases and eventually becomes zero. Also as proposed, contribution of tandem data to the accuracy can clearly be seen since for each model (whether regular MSHMM or stream-tied MSHMMs) tandem stream employed results are better from audio-visual results without tandem data at almost every SNR level. The increase in accuracy achieved by visual tandem streams evidence the improvement of the proposed method over conventional observation only based audio-visual speech recognisers, which are state of the art in audio-visual speech recognition.

Comparing results across different models can give information about using regular MSHMM or two different training strategies for stream-tied MSHMM. The regular MSHMM model is used as an initial model to generate stream-tied models and since state coupling, stream tying and parameter (whether only transition or both transition and emission probabilities) training changes the structure of the models, the results do differ for audio-only and video-only columns across models. The increase in the accuracy for tandem employed models between regular and stream-tied MSHMM trained with first method shows the benefit of taking asynchrony into consideration. The increase in the accuracy for tandem employed models at lower SNR values between stream-tied MSHMM trained with two methods shows that training emission parameters together with transition probabilites increase accuracy when weights of the video based streams are higher than the audio stream. Curiously, for higher SNR

Table 1. Best results using synchronous MSHMM and CHMM implemented as stream-tied MSHMM.

	Synchronous MSHMM				CHMM (Trained with Method 1)				CHMM (Trained with Method 2)			
SNR	**Audio**	**Video**	**Audio Visual**	**AV and Tandem**	**Audio**	**Video**	**Audio Visual**	**AV and Tandem**	**Audio**	**Video**	**Audio Visual**	**AV and Tandem**
Clean	100	36.67	100	**100**	100	35.56	100	**100**	100	36.94	100	**100**
20	99.17	36.67	100	**100**	99.72	35.56	99.72	**99.72**	97.22	36.94	98.61	**99.72**
15	93.61	36.67	96.67	**96.67**	92.50	35.56	96.67	**96.67**	85.55	36.94	93.61	**93.61**
10	74.44	36.67	81.67	**85.00**	74.17	35.56	82.22	**89.44**	60.28	36.94	75.55	**80.83**
5	37.50	36.67	54.44	**62.50**	36.67	35.56	49.44	**66.11**	31.11	36.94	43.89	**67.22**
0	11.39	36.67	36.67	**52.78**	11.94	35.56	36.94	**58.33**	17.78	36.94	39.44	**56.39**
-5	9.44	36.67	36.67	**46.39**	10.00	35.56	35.56	**54.44**	10.56	36.94	36.94	**54.72**
-10	6.11	36.67	36.67	**45.56**	6.11	35.56	35.56	**48.33**	8.33	36.94	36.94	**54.72**
-15	2.78	36.67	36.67	**45.56**	3.33	35.56	35.56	**48.33**	9.72	36.94	36.94	**54.72**
-20	6.94	36.67	36.67	**45.56**	6.94	35.56	35.56	**48.33**	7.5	36.94	36.94	**54.72**

values, the transition-only update method seems to work better. In our future studies, we will work on developing training strategies that will work best for all SNR values.

7. CONCLUSION

We presented a new method for audio-visual speech recognition which uses multiple visual tandem features in parallel with regular audio and video features in a multi-stream HMM framework. We experimented with synchronous and asynchronous HMM methods using multi-stream HMM and coupled HMM. It is shown that using visual tandem features improve recognition accuracy for high and low SNR values. In addition, asynchrony modeling greatly improves accuracy in low SNR conditions. We believe this method is an attractive approach in audio-visual speech recognition and there are many potential areas for improving the method such as using different classifiers, utilizing an increased number of tandem streams and employing better initilization and training methods for the coupled HMM which we plan to pursue as future work.

Table 2. Stream weights for the results in the last column of Table 1.

SNR	Audio	Video	NN	SVM
Clean	0.75	0.25	0	0
20	0.75	0	0	0.25
15	0.75	0.25	0	0
10	0.5	0.25	0	0.25
5	0.5	0	0.25	0.25
0	0.25	0	0.25	0.5
-5	0	0.75	0.25	0
-10	0	0.75	0.25	0
-15	0	0.75	0.25	0
-20	0	0.75	0.25	0

8. REFERENCES

[1] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, Jan 1986.

[2] Hynek Hermansky, Daniel P. W. Ellis, and Sangita Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *ICASSP*, 2000.

[3] S. Dupont and J. Luettin, "Audio-visual speech modelling for continuous speech recognition," *IEEE Transactions on Multimedia*, 2000.

[4] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1274–1288, 2002.

[5] Guoyun Lv, Dongmei Jiang, Rongchun Zhao, and Yunshu Hou, "Multi-stream asynchrony modeling for audio-visual speech recognition," in *International Symposium on Multimedia*, Washington, DC, USA, 2007, pp. 37–44, IEEE Computer Society.

[6] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W. Senior, "Recent advances in the automatic recognition of audio-visual speech," in *PROC. IEEE*, 2003, pp. 1306–1326.

[7] Aravind Ganapathiraju, *Support vector machines for speech recognition*, Ph.D. thesis, Mississippi State, MS, USA, 2002, Professor - Picone, Joseph.

[8] T.S. Chu, S.M. Huang, "Audio-visual speech modeling using coupled hidden markov models," in *ICASSP*, 2002, vol. 2, pp. 2009–2012.

[9] Stephane Pigeon and Luc Vandendorpe, "The m2vts multimodal face database (release 1.00).," in *AVBPA*. 1997, vol. 1206 of *Lecture Notes in Computer Science*, pp. 403–409, Springer.

[10] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*, pp. 374–388, 1976.

[11] H. Karabalkan, H. Erdogan, "Information fusion techniques in audio-visual speech recognition," in *Signal Processing and Communications Applications Conference*, 2009, pp. 504–507.

[12] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 214–259, 1992.

[13] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition.," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[14] Carnegie Mellon University, "Noisex-92 database," `http://www.speech.cs.cmu.edu/comp.speech/Section1/-Data/noisex.html`, 2010.

[15] ITU, "G.191 : Software tools for speech and audio coding standardization," `http://www.itu.int/rec/T-REC-G.191/en`, 2010.