Cross database training of audio-visual hidden Markov models for phone recognition

Shahram Kalantari, David Dean, Houman Ghaemmaghami, Sridha Sridharan, Clinton Fookes

Speech, Audio, Image, and Video Technologies, Science and Engineering Faculty
Queensland University of Technology, 2 George St, Brisbane, Australia 4000
{s1.kalantari, h.ghaemmaghami, s.sridharan, c.fookes}@qut.edu.au, ddean@ieee.org

Abstract

Speech recognition can be improved by using visual information in the form of lip movements of the speaker in addition to audio information. To date, state-of-the-art techniques for audio-visual speech recognition continue to use audio and visual data of the same database for training their models. In this paper, we present a new approach to make use of one modality of an external dataset in addition to a given audio-visual dataset. By so doing, it is possible to create more powerful models from other extensive audio-only databases and adapt them on our comparatively smaller multi-stream databases. Results show that the presented approach outperforms the widely adopted synchronous hidden Markov models (HMM) trained jointly on audio and visual data of a given audio-visual database for phone recognition by 29% relative. It also outperforms the external audio models trained on extensive external audio datasets and also internal audio models by 5.5% and 46% relative respectively. We also show that the proposed approach is beneficial in noisy environments where the audio source is affected by the environmental noise.

Index Terms: Phone recognition, synchronous hidden Markov model, fused hmm adaptation, cross database training.

1. Introduction

Automatic human speech recognition has been an ongoing research area since the 1950s [1], as it was clear from human studies that most of the speech information is contained in audio signals. Many commercial speech recognition applications have been developed in this area. However, complete speech recognition systems under different situations, especially in noisy conditions and open vocabulary systems, is still a difficult ongoing research problem.

Human speech perception is bimodal in nature: humans combine audio and visual information in deciding what has been spoken, especially in noisy environments. The visual modality benefit to speech intelligibility in noise has been quantified as far back as in the research of Sumby et al. in 1954 [2]. The first attempt to achieve better recognition performance by using visual information in addition to the audio modality was in 1989 by Yuhas et al. [3]. In this work, pixel values of the lip region were fed into a neural network to estimate the acoustic spectrum. This estimated spectrum was then combined with the true spectrum and used as input to an acoustic vowel recognizer. The results showed that this system outperforms the acoustic only recognizer. This pioneering work opened the door to significant further research which examined numerous approaches to combine and integrate audio and visual sources, particularly using the HMM [4, 5, 6], most of which proved that using visual information in addition to the audio data improves the task of speech recognition, especially in noisy conditions.

Generally, for combining audio and visual sources there are three integration strategies [4]:

- Early integration, which involves a concatenation of feature vectors of audio and visual modalities or a transformation of these feature vectors using PCA or LDA methods before classification. The advantage of this method is its simplicity, however, it cannot capture the reliability of each modality.
- Middle integration, in which covering classifiers are designed in a way that are able to handle features from different modalities.
- Late integration, where scores or decisions of individual classifiers for each modality are combined before reaching the final decision.

For audio-visual speech processing all three levels of fusion have been considered. Synchronous multi-stream HMM (SHMM) is one of the proposed models for middle integration strategy which can be viewed as a regular single stream HMM, but with two observation emission Gaussian mixture models (GMMs) for each of the two streams in each state [4]. Brand et al. [7] also proposed coupled multi-stream HMM as a middle integration strategy in which all states of an audio HMM are connected to all states of a visual HMM and then the coupled HMM is jointly trained with training data. Training a coupled HMM is a difficult issue [8], as it needs to learn more parameters. In addition, this problem is exacerbated by the current limited audio visual databases.

Training a SHMM could be done by training two separate single stream HMMs independently and then combining them together. Alternatively, the entire SHMM could be trained jointly. Fused HMM (FHMM) adaptation was introduced by Pan et al. [9] as an alternative modelling method to maximize the mutual information between the two modalities. In this method, first a HMM is trained solely on the dominant modality. For example, for the task of speech recognition, audio information is more important than visual information. Therefore, in the first step a HMM is trained based on the audio data. In the next step, a discrete vector-quantisation classifier for the subordinate modality is combined within each state. Inspired from this method, Dean et al. [8] proposed fused HMM adaptation as a new way of multi-stream HMM training for the task of speech recognition. In this approach, first a HMM model is trained based on audio information. Then the best hidden state sequence of the acoustic HMM is determined over the training data and then for each state of the acoustic HMM, a visual Gaussian mixture model (GMM) is trained based upon the visual observations that coincide with the acoustic state in the training

data. This visual GMM then will be appended to the already existing acoustic GMM in each state to produce an audio-visual synchronous HMM.

In order to jointly train an audio-visual HMM, a fairly large and annotated audio-visual database is needed which is limited due to capturing difficulties, annotation cost, and time limitations. As the initial audio model in fused HMM adaptation could be trained independently, in this paper we propose the idea of cross database training to make use of existing powerful audio models which can be trained on external audio databases. This approach enables us to make use of other existing audio-only databases to train more accurate audio models. We compare our proposed approach with a widely adopted jointly trained two-stream SHMM for the task of phone recognition. In addition, we report the results using audio HMMs trained on solely external and also internal audio and internal visual data in clean and noisy environments.

2. Fused HMM adaptation

The Fused HMM was first introduced by Pan et al. [9] as a multi-stream modelling method for audio-visual speaker recognition. In this approach, first a continuous HMM is created for the dominant modality. Then a discrete vector-quantization classifier for the subordinate modality is created within each state and finally combined with the initial HMM. Training of the subordinate classifier is based on forced-alignment of the initial HMM on the training set. Dean et al. [8] extended this work by using a continuous classifier for the subordinate modality instead of a discrete one. More specifically, their fused HMM adaptation for the task of speech recognition consists of the following four steps:

- Train a single stream HMM based on audio data.
- Find the best hidden state alignment of the audio HMM by force-aligning the training transcriptions on each audio observation.
- Using the results of forced-alignment, train a visual GMM for each state based upon the visual observations which line up with best hidden state alignment.
- Append the visual GMMs to corresponding states and create audio-visual HMMs.

In Dean et al's. work [8], training data was limited to one audio-visual dataset. However, training of the audio models is independent of training the visual GMMs. Moreover, while there is quite a large number of huge annotated audio datasets such as TIMIT, WSJ, Switchboard, and so on, currently there are not many large, publicly available audio-visual datasets that are completely annotated and suitable for speech recognition. In this work, we propose using a more powerful audio model which could be trained on external audio datasets and adapting it to the visual observations of our audio-visual dataset using the FHMM framework. This approach enables us to benefit from external audio models which can potentially have much more accuracy and thus enable improved audio-visual speech recognition performance.

3. Audio-visual phone recognition

3.1. Phone recognition task

Audio-visual phone recognition is defined as recognizing the sequence of phones of a multimedia document. Phone recognition does not require any vocabulary and is language independent as well. Such systems are more popular for open-vocabulary keyword spotting purposes [10, 11] which could potentially search for any sequence of phones. However, the accuracy of such systems is rather lower compared with corresponding word recognition systems [12]. Phone recognition accuracy is reported for each experiment in this study and is defined as,

$$\frac{(H-I)}{N} * 100\%, \tag{1}$$

where H is the number of correctly recognized phones, I is the number of incorrectly inserted phones, and N is the total number of actual phones in the ground truth.

Table 1: Two configurations used for training, tuning, and test

	Config 1	**Config 2**
Train	F02, F04, F06, F08 F10, F11, M01, M03 (CID)	F02, F04, F06, F08 F10, F11, M01, M03 (TIMIT)
Tune	F03, F07, M02 (CID)	F03, F07, M02 (TIMIT)
Test	1) F05, F09, M04 (CID) 2) F05, F09, M04 (TIMIT)	1) F05, F09, M04 (TIMIT) 2) F05, F09, M04 (CID)

3.2. Training and testing datasets

Training, tuning, and testing data were extracted from the audio-visual database of spoken American English (AVDBAE) [13]. There are 14 speakers in this database including 10 females (F02-F011) and 4 males (M01-M04). Each participant reads 238 different words and 166 different sentences. The spoken text were drawn from the following sources:

- Central Institute for the Deaf (CID) Everyday Sentences (Lists A-J)
- Northwestern University Auditory Test No. 6 (Lists I-IV)
- Vowels in /hVd/ context (separate words)
- Texas Instruments/Massachusetts Institute for Technology (TIMIT) sentences

In this work, we only used the CID and TIMIT sentences which are quite longer than the other utterances. These sentences were divided into different portions to be used for training, tuning, and testing purposes. We also defined two configurations for our experiments: in the first configuration, CID sentences are used for training and tuning and in the second configuration, TIMIT sentences are used for training and tuning. We report the testing result on both sets of TIMIT and CID sentences. This was done to investigate the effect of using different and the same set of sentences in training and testing. Table 1 summarizes the portions of the dataset utilised and the two configurations used in this work. For training external audio models, TIMIT, WSJ1, and 160 hours of speech from Switchboard-1 Release 2 was used.

3.3. Feature extraction

Perceptual linear prediction (PLP) based cepstral features were extracted to represent the acoustic features in our experiments. Each feature vector consisted of the first 13 PLPs including the zeroth, as well as the first and second time derivatives of those

13 features. These 39 dimensional feature vectors were extracted from every 10 milliseconds of 25-millisecond windows using Hamming-windowed speech signals.

In order to extract visual features, the Fourier Lucas-Kanade algorithm proposed by Lucey et. al [14] was used to extract the lip region-of-interest (ROI) from 29.97 fps video data. This method has been shown to work better than the Viola-Jones algorithm in a semi automatic manner [15]. After ROI extraction, the mean ROI over the video segment is removed. In the next step, a two-dimensional discrete cosine transform (DCT) is applied to the mean-removed ROI, with the 100 top DCT coefficients according to the zig-zag pattern retained, resulting in a 'static' visual feature vector. Finally, in order to extract dynamic speech information, 7 neighbouring adjacent feature vectors centred at the current vector were passed to inter-frame linear discriminant analysis (LDA) to achieve a 60 dimensional LDA feature vector.

3.4. Baseline systems

In addition to our cross database trained fused HMM adapted models (CDB-EX-FHMM system), the following four baseline systems are also trained and tested for comparison:

- Internal audio HMMs (IAHMM system) which are trained using audio data of our audio-visual database.
- External audio-only models (EAHMM system) which are trained using audio data of a extensive external audio. database.
- Visual models (VHMM) trained on lip images of our audio-visual database.
- Jointly trained widely adopted two-stream HMM (JAVHMM system) trained using audio-visual data of our audio-visual database.

All baseline systems were trained and tested with the HTK Toolkit [16]. In all systems, a bi-gram word language model is used. Phonetic decoding is performed by decoding a lattice of words, then expanding these tokens into their corresponding sequences of phones using a pronunciation lexicon, whilst maintaining lattice structure. This approach has shown to perform better than using phone language models for phone recognition [17]. HMM parameters including the number of states and mixtures, grammar scale, insertion penalty, number of tokens, as well as stream weights in case of audio-visual HMMs were tuned on the tuning set after training. The best set of tuned parameters were then selected to report the accuracy on the test set.

As the sampling rate of visual frames and audio frames were not equal, for training and testing of audio-visual systems, the closest visual feature vector to each audio feature vector was selected and appended to it to create a single 99 dimensional feature vector.

3.5. Fused HMM adaptation with cross database training

Our method of cross-database fused HMM adaptation consists of the following steps:

1. Train an audio HMM for each phone on an external large audio database. In this paper, a monophone HMM is trained for each phoneme class, which is a 32 mixture monophone HMM, with 3 emitting states.
2. For each audio observation in the given audio-visual database, find the best sequence of states of the audio HMM corresponding to that audio segment by forced-alignment.
3. Train a global background GMM using all visual feature vectors of the audio-visual database.
4. For each state of each model, adapt the background model to the corresponding visual feature vectors resulted from step 2 and append it to that state as the visual GMM.

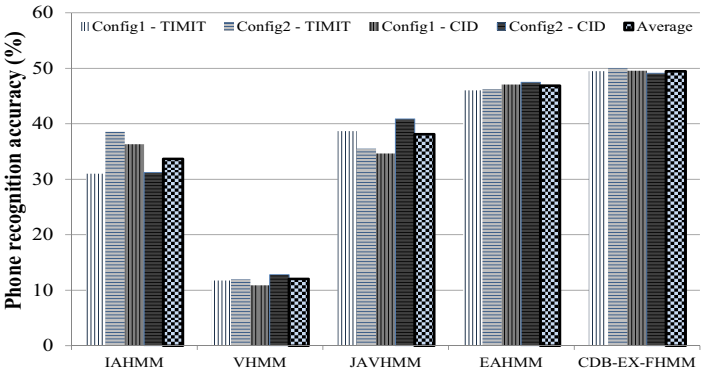

Figure 1: Phone recognition accuracy for different systems using different configurations.

The number of visual GMMs as well as other parameters mentioned for other baseline systems are tuned on the tuning set. It should be noted that the best weights for the JAVHMM were 0.91/0.09 and 0.86/0.14 when tuning on Config 1 and Config 2 tuning data respectively for audio/visual modality. These weights for the CDB-EX-FHMM were: 0.93/0.07 and 0.85/0.15 respectively.

4. Results and discussion

The results of the phone recognition experiments on the testing sets are shown in Figure 1. From the phone recognition accuracy bars in this figure, it can be seen that as expected, using external audio models trained on an external large corpus generally improves speech recognition compared with internal audio models which are trained on audio data of the given audio-visual database.

When IAHMM is trained on TIMIT sentences and also tested on the TIMIT test sentences, the accuracy of EAHMM is about 10% absolute higher than that of IAHMM. However, when it is tested on CID test sentences, we can see that EAHMM outperforms IAHMM by about 16% absolute. IAHMM is trained on a subset of AVDBAE which is limited to less than an hour of speech for each of the 8 speakers in the training set.In contrast, EAHMM is trained on TIMIT, WSJ1, and 160 hours of speech from SWB databases. This proves that large amounts of audio data could make a more generalized model for recognition even when training data is selected completely independent of the testing set.

Moreover, the phone recognition accuracy in the JAVHMM system is more than that of the IAHMM system. Phone recognition in the VHMM has the worst accuracy which is around 12% in average. This confirms that, as it has been shown before, incorporating visual information in addition to audio information within the SHMM framework is more accurate than internal audio or visual models. Although this is not a huge improvement, it suggests that phone recognition could still benefit from the visual modality.

As illustrated in Figure 1, the proposed approach outperforms all other systems. The average of phone recognition

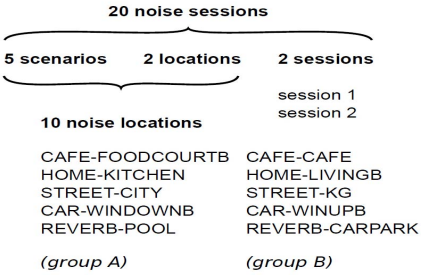

Figure 2: An overview of the structure of the QUT-NOISE-TIMIT corpus of noisy speech recordings [18]

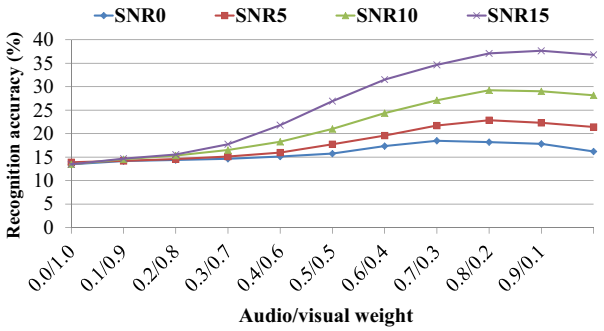

Figure 3: The average phone recognition accuracy of different audio visual models across different weights of audio/visual streams in different SNR levels of noise

accuracy in the CDB-EX-FHMM system is about 29%, 5.5% and 46% relative higher than that of the JAVHMM, EAHMM and IAHMM respectively. In fact, CDB-EX-FHMM benefits from advantages of both EAHMM and JAVHMM systems at the same time. First it uses the EAHMM directly to create audio models and then it adapts them to visual data of the testing database. These two stages of training are independent. Therefore it provides a framework for using different data in each stage and as a result, it is able to use any well-trained model for the first stage and then it adapts it to the visual data of the testing database. Particularly, because of limited large and publicly available annotated audio-visual datasets, this approach is very beneficial for audio-visual speech processing tasks.

To investigate the robustness of the proposed approach, we performed phone recognition under different noise conditions. We added noise to the audio source of the video recording with various signal to noise ratios (SNR) and performed phone recognition experiments with different audio-visual weights. We utilised the QUT-NOISE database [18] for this purpose which is designed for performance evaluation of speech processing algorithms across different levels of noise recordings. In this dataset 5 noise scenarios were considered, where for each scenario, two locations were used for recording environment noise. Two sessions were conducted for recording the noise in each location which resulted in total of 20 noise recordings. Figure 2 demonstrates the structure of the QUT-NOISE corpus.

For the experiments of this study, we utilized the noise recordings of home-living environment of the QUT-NOISE database. The reason is that one of the applications of audio visual speech recognition systems is for video chats which usually take place in home environments where there is TV noise, kitchen noise, and etc. The noise data are added to AV-DB-AM-ENG recordings with different SNR levels. We used the first session of each recording for tuning the model parameters and the second session is used for testing. Figure 4 shows the average of phone recognition performance of the proposed audio-visual systems across different weights of audio/visual streams on test data in different SNR levels of noise.

We tried various audio/visual weights from 0.0/1.0 to 1.0/0.0 in steps of 0.1 and reported the results of phone recognition accuracy on the test data. Figure 3 shows that the proposed approach for phone recognition benefits from visual information in different noise environments. The audio/visual weight value of 0.0/1.0 indicates the visual only phone recognition and value of 1.0/0.0 denotes the audio only phone recognition. In general, the performance of phone recognition increases from audio/visual weight of 0.0/1.0 (visual only) by increasing audio weight until it reaches an optimum point and then it gets slightly decreased towards the weight of 1.0/0.0 (audio only). As it can be seen, in environments with high noise levels (SNR0), the optimum audio/visual weight value is around 0.7/0.3, while in low noise levels (SNR15) it happens around 0.9/0.1. This proves that using visual information in addition to audio data is more beneficial in noisy environments where the audio source is contaminated and visual source can help to resolve dis ambiguities. It is also worth mentioning that the results of the cross database training technique is better than either individual modality ones at all noise levels.

5. Conclusion

In this work, we showed that fused HMM adaptation could be used with more powerful external audio models to improve the video phone recognition compared with the original fused HMM adaptation and widely adopted jointly trained audio-visual HMMs. Fused HMM adaptation consists of independent stages, the first of which is training an audio model and doing forced-alignment of the video frames. Training the audio models could be performed on external databases to benefit from large-size audio databases and create more accurate phone models. These models, consequently, provide better alignment for video frames which then create visual GMMs to be appended to primary audio models. This approach, not only provides better audio models for final audio-visual models, it will provide better alignment compared with the original fused HMM adaptation approach which in fact provides better visual GMMs to be appended to audio models. We also showed that incorporating visual information provides robustness to the phone recognitions system within the proposed framework. Specially, using the proposed approach were shown to be more beneficial in highly noisy environments.

In this work, we didn't consider training the state transition probabilities of the final audio visual HMM. As a future work, it could be investigated if re-training the final model to update the state-transition probabilities can help the performance of phone recognition. This could be probably done by adapting the external audio models on the audio frames of the given audio-visual database in the first step and then adapting them to the visual frames of the audio-visual database improves the phone recognition accuracy. As another solution, the proposed model could be considered as an initial model for Baum-Welch re-estimation algorithm to update the whole model parameters.

6. Acknowledgements

This work has been supported by the Australian Cooperative Research Center for Smart Services.

7. References

[1] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 5, pp. 1060–1089, May 2013.

[2] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," vol. 26, no. 2, pp. 212–215, Mar. 1954. [Online]. Available: http://dx.doi.org/10.1121/1.1907309

[3] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine*, 1989.

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

[5] L. J. Rothkrantz, J. Wojde, and P. Wiggers, "Fusing data streams in continuous audio-visual speech recognition," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, V. Matouek, P. Mautner, and T. Pavelka, Eds. Springer Berlin Heidelberg, 2005, vol. 3658, pp. 33–44.

[6] S. Chin, K. Seng, and L.-M. Ang, "Audio-visual speech processing for human computer interaction," in *Advances in Robotics and Virtual Reality*, ser. Intelligent Systems Reference Library, T. Gulrez and A. Hassanien, Eds. Springer Berlin Heidelberg, 2012, vol. 26, pp. 135–165.

[7] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden markov models for complex action recognition," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, Jun 1997, pp. 994–999.

[8] D. B. Dean, P. J. Lucey, S. Sridharan, and T. J. Wark, "Fused HMM-adaptation of multi-stream HMMs for audio-visual speech recognition," in *8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*. Antwerp: International Speech Communication Association (ISCA), 2007, pp. 666–669. [Online]. Available: http://eprints.qut.edu.au/13351/

[9] H. Pan, S. Levinson, T. Huang, and Z.-P. Liang, "A fused hidden Markov model with application to bimodal speech processing," *Signal Processing, IEEE Transactions on*, vol. 52, no. 3, pp. 573–581, March 2004.

[10] S. Kalantari, D. Dean, and S. Sridharan, "Topic dependent language modelling for spoken term detection," in *Europian Signal Processing Conference, 2014. Proceedings. (EUSIPCO 2014).*, 2014. [Online]. Available: http://eprints.qut.edu.au/75760/

[11] S. Kalantari, D. B. Dean, and S. Sridharan, "Phonetic spoken term search using topic information," in *Science and Speech Technology conference*, 2014.

[12] R. G. Wallace, "Fast and accurate phonetic spoken term detection," Ph.D. dissertation, Queensland University of Technology, 2010. [Online]. Available: http://eprints.qut.edu.au/39610/

[13] S. Richie, Carolyn Warburton and M. Carter, "Audiovisual database of spoken American English," *Linguistic Data Consortium,*, 2009.

[14] S. Lucey, R. Navarathna, A. B. Ashraf, and S. Sridharan, "Fourier Lucas-Kanade algorithm," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 6, pp. 1383–1396, June 2013.

[15] S. Kalantari, R. Navarathna, D. B. Dean, and S.Sridharan, "Visual front-end wars : Viola-Jones face detector vs Fourier Lucas-Kanade," in *International Conference on Auditory Visual Speech Processing 2013*, B. Denis and B. Jonas, Eds., Ternélia resort Le Pré du Lac, Annecy, France, 2013. [Online]. Available: http://eprints.qut.edu.au/62749/

[16] S. Young, G. Everman, T. Hain, D. Kershaw, G. Moore, J. Odell, V. V. Ollason, D. D. Povey, and P. Woodland, "The htk book (for htk version 3.2.1)," 2002.

[17] R. Wallace, B. Baker, R. Vogt, and S. Sridharan, "The effect of language models on phonetic decoding for spoken term detection," in *ACM Multimedia Workshop on Searching Spontaneous Conversational Speech*, 2009, pp. 31–36.

[18] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The qut-noise-timit corpus for the evaluation of voice activity detection algorithms," in *Interspeech 2010*, Makuhari Messe International Convention Complex, Makuhari, Japan, September 2010, to download the full database, visit: https://wiki.qut.edu.au/display/saivt/QUT-NOISE-TIMIT. [Online]. Available: http://eprints.qut.edu.au/38144/