

QUT Digital Repository:  
<http://eprints.qut.edu.au>



Dean, David B. and Lucey, Patrick J. and Sridharan, Sridha and Wark, Timothy J. (2007) Fused HMM-adaptation of multi-stream HMMs for audio-visual speech recognition. In *Proceedings Interspeech 2007*, pages pp. 666-669, Antwerp.

Power Point Presentation

© Copyright 2007 (The authors)

# Fused HMM-Adaptation of Multi-Stream HMMs for Audio-Visual Speech Recognition

David Dean\*, Patrick Lucey\*, Sridha Sridharan\*,  
and Tim Wark\*\*

Presented by David Dean

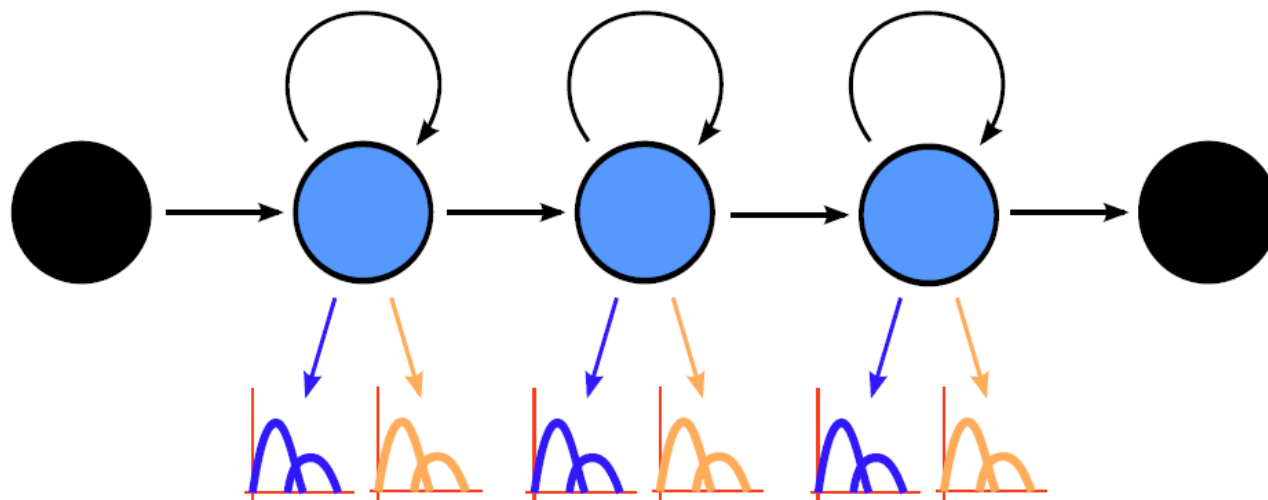
*Slides will be available at <http://www.davidbdean.com/category/publications>*

# Why audio-visual speech recognition?

- Why speech?
  - Speech is still not a natural method of human-computer interaction
  - Many problems come about due to use in poor or loosely controlled conditions
- Why audio-visual?
  - Human speech is inherently bimodal in nature
  - Video can be used to complement acoustic speech
  - Noise affects each modality differently

# Synchronous HMMs (SHMMs)

- Provide a middle ground between feature-fusion and asynchronous HMMs (AHMMs)
- Can model reliability of streams independently
- Are simpler to train and implement than AHMMs

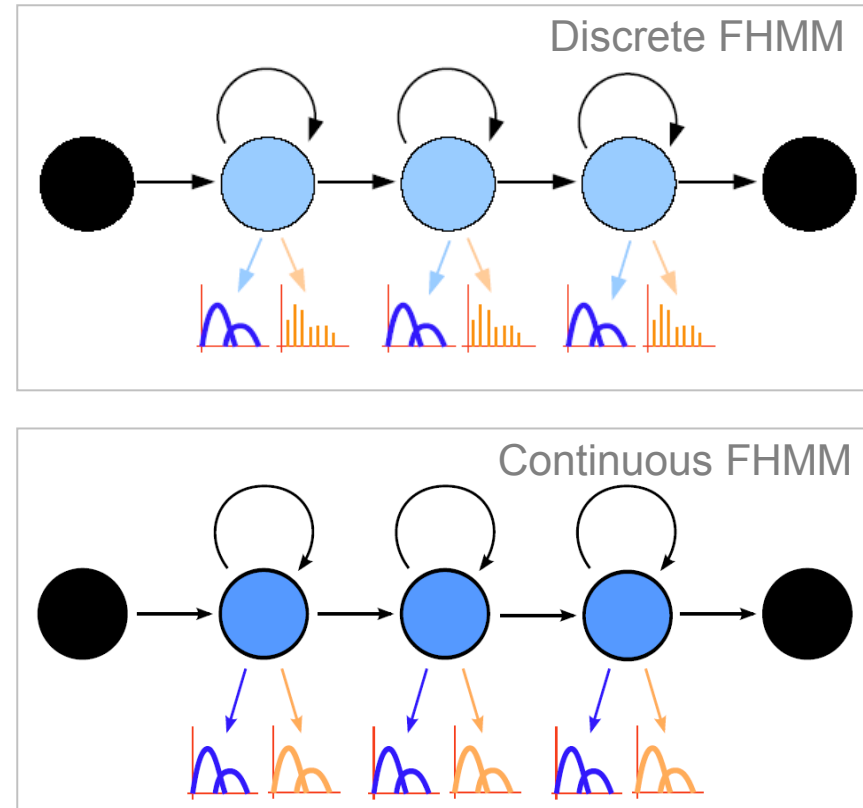


# SHMM Training

- *Neti et al (2000)* suggest two methods
  - Combine two independently trained HMMs (separate training), or
  - Train the two streams simultaneously (joint-training)
- Joint-training performs better because it ensures that the emission models are kept synchronous
- Fused HMM-Adaptation is a novel method of SHMM training based upon Fused HMMs

# Fused HMMs

- Originally introduced for audio visual *speaker* recognition by Pan et al (2004)
- Video models trained on audio alignment (or visa-versa)
- The original implementation treated the secondary domain as discrete
- Our continuous FHMMs (Dean 2006) model both modalities with GMMs (ie. SHMM)

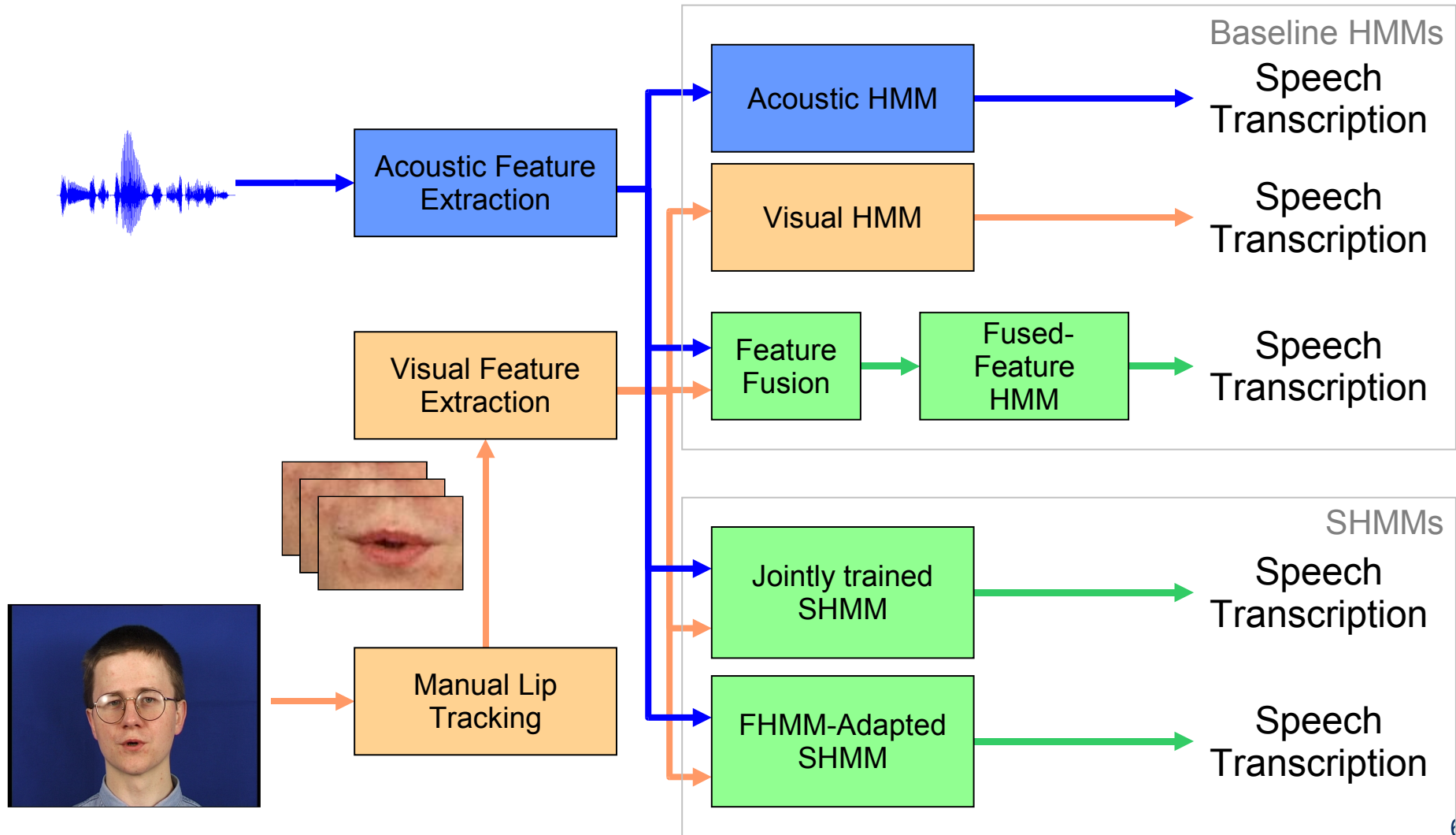


H. Pan, S. Levinson, T. Huang, and Z.-P. Liang, "A fused hidden markov model with application to bimodal speech processing," *IEEE Transactions on Signal Processing*, vol. 52, no. 3, pp. 573–581, 2004.

D. Dean, S. Sridharan, and T. Wark, "Audio-visual speaker verification using continuous fused HMMs," in *VisHCI 2006*, 2006.

5

# Experimental setup



# Training and testing datasets

- Training and testing configuration was based on the XM2VTSDB speaker recognition protocol (Messerb et al. 1999)
- 295 speakers
  - 200 'client' speakers used to train speech models
  - 95 'imposter' speakers used to test
- 4 sessions per speaker
  - Train with first two sessions
  - Evaluate with third
  - Test with fourth

K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Audio and Video-based Biometric Person Authentication (AVBPA '99)*, Washington D.C., 1999, pp. 72–77.

7

# Feature extraction

- Audio
  - PLP – 12 + 1 energy, + deltas and accelerations = 43 features
- Video
  - Lip ROI manually tracked every 50 frames
  - Hierarchical LDA based video feature extraction (Potamianos et al 2003)
    - Mean image removed to remove speaker-specific information
    - DCT performed on mean-removed images
    - Adjacent DCT frames concatenated and LDA'd based on speech events

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

8

# Synchronising Audio and Visual Features

- Some of the experiments performed require audio-visual data to be synchronous
  - Feature Fusion
  - Jointly-trained and FHMM-Adapted SHMM
- Synchronisation is performed by choosing the closest video feature to a particular audio feature
  - No interpolated estimation is performed

# HMM Training

- All HMMs (including the FHMM-adapted) were trained using 13 states and 10 mixtures
- All HMMs (except the FHMM-adapted) were trained with the HTK Toolkit (Young et al 2002)
- SHMMs were weighted based on an audio weight of  $\alpha$  and a video weight of  $(1-\alpha)$ 
  - $\alpha = 0.9$  was found to be best for jointly-trained SHMMs

# FHMM-Adaptation

- FHMM-adaptation adapts the video state-models from an existing audio HMM to generate a SHMM
  1. For each audio observation, find the best hidden state alignment of the audio HMM
  2. Train video state-models based on the hidden state alignments from (1)

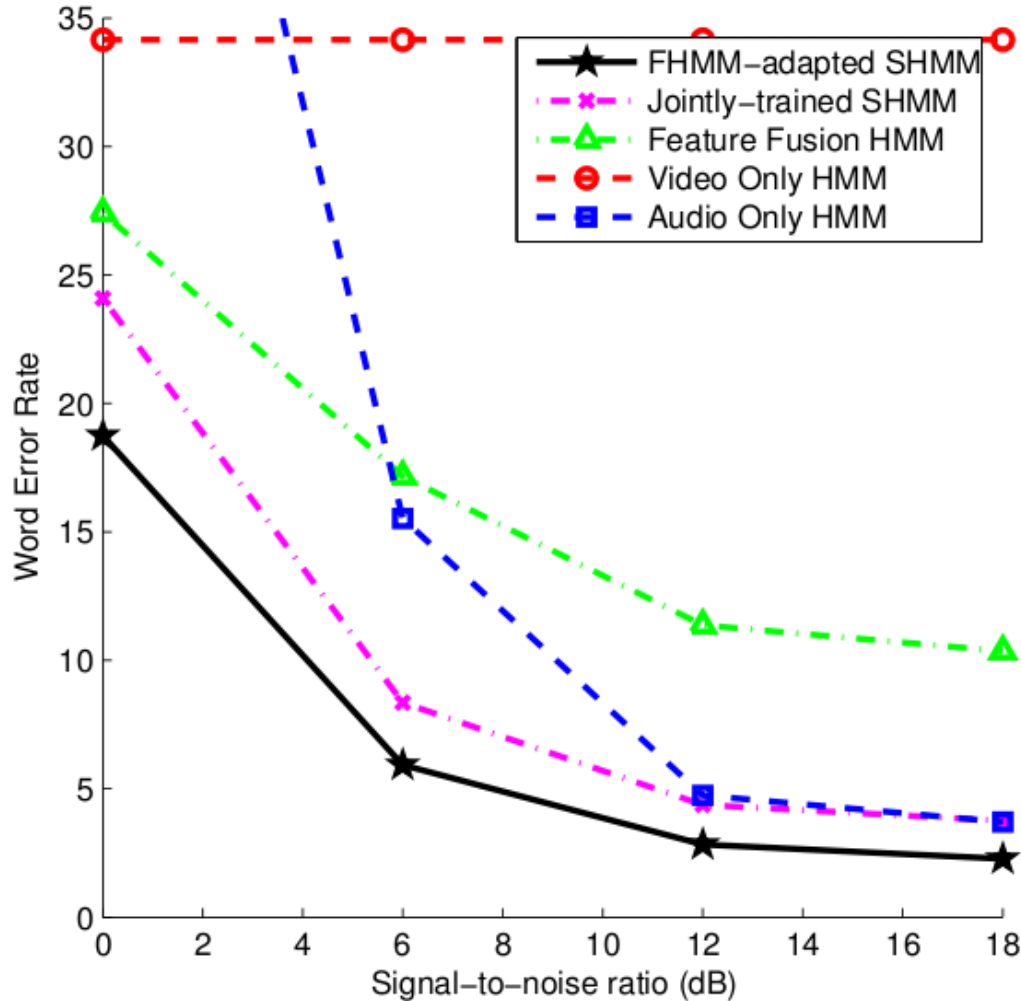
# SHMM Testing

- The jointly-trained and FHMM-adapted SHMMs were tested in an identical manner
- Scores from each streams model were normalised within the SHMM state
  - Essentially pre-weighting the streams such that the output-score variances are equal
  - See our AVSP 2007 (Dean et al, 2007) paper for more detail
- Post-normalisation testing  $\alpha$  of 0.5 was found to give best performance

D. Dean, P. Lucey, S. Sridharan, and T. Wark, "Weighting and normalisation of synchronous HMMs for audio-visual speech recognition," in *Auditory-Visual Speech Processing 2007*, 2007

12

# Speech Recognition Results



- SHMMs can represent each stream independently for improved performance over feature fusion
- Fused-HMM adaptation improves WER at all noise levels

# Examining Audio Speech Recognition

- Extract individual modality performance from SHMMs with  $\alpha = 0$  and  $\alpha = 1$
- FHMM audio performance is basically the same as baseline audio HMM (by definition)
- Jointly-trained audio performance is degraded, particularly in noisy environments

| Audio Model     | Audio-only WER over noise |       |      |      |
|-----------------|---------------------------|-------|------|------|
|                 | 0dB SNR                   | 6dB   | 12dB | 18dB |
| FHMM-adapted    | 63.98                     | 15.41 | 5.16 | 3.82 |
| Audio-only HMM  | 64.12                     | 15.52 | 4.75 | 3.71 |
| Jointly-trained | 68.41                     | 18.82 | 4.75 | 3.79 |

# Examining Video Speech Recognition

- Video performance of FHMM-adapted SHMM is superior to both
- Video performance of jointly-trained SHMM is significantly degraded to the baseline video HMM
- This video speech recognition performance is the main reason for the FHMM-adapted SHMMs improvement over jointly trained

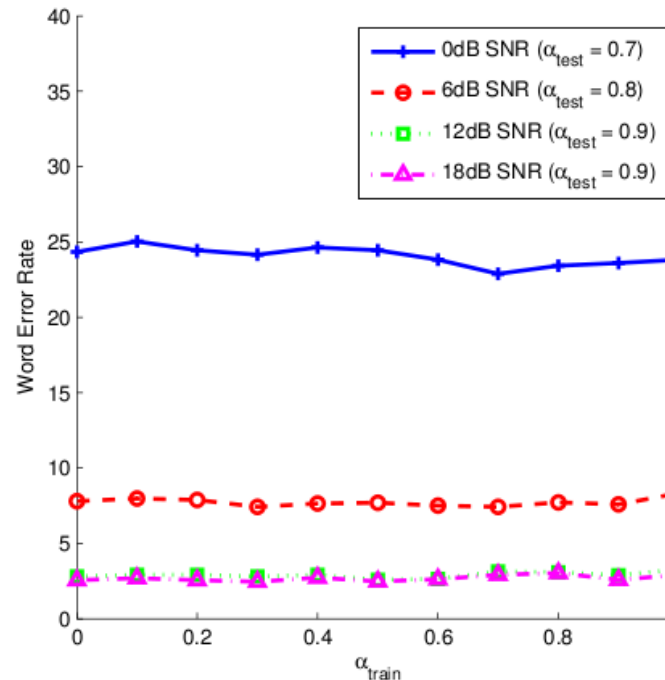
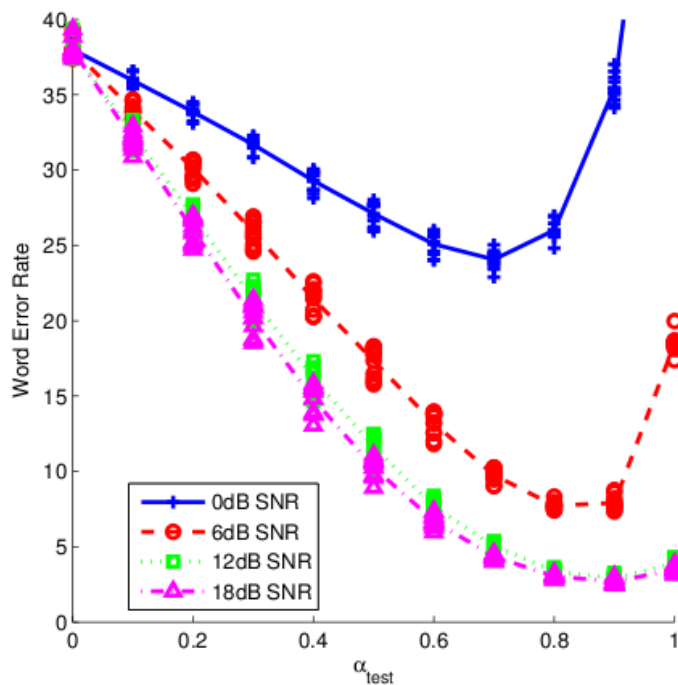
| Video Model     | Video-only WER |
|-----------------|----------------|
| FHMM-adapted    | 30.80          |
| Video-only HMM  | 34.15          |
| Jointly-trained | 40.63          |

# Conclusion

- By using audio alignments to train the video models, FHMM-adapted SHMMs improve performance over jointly-trained
- Mostly due to improved video speech recognition
  - Greatest WER improvement in noisy audio where video is more important
- However, this is *not* because the audio alignment is better than the jointly trained alignment
  - This is continuing research, not in paper

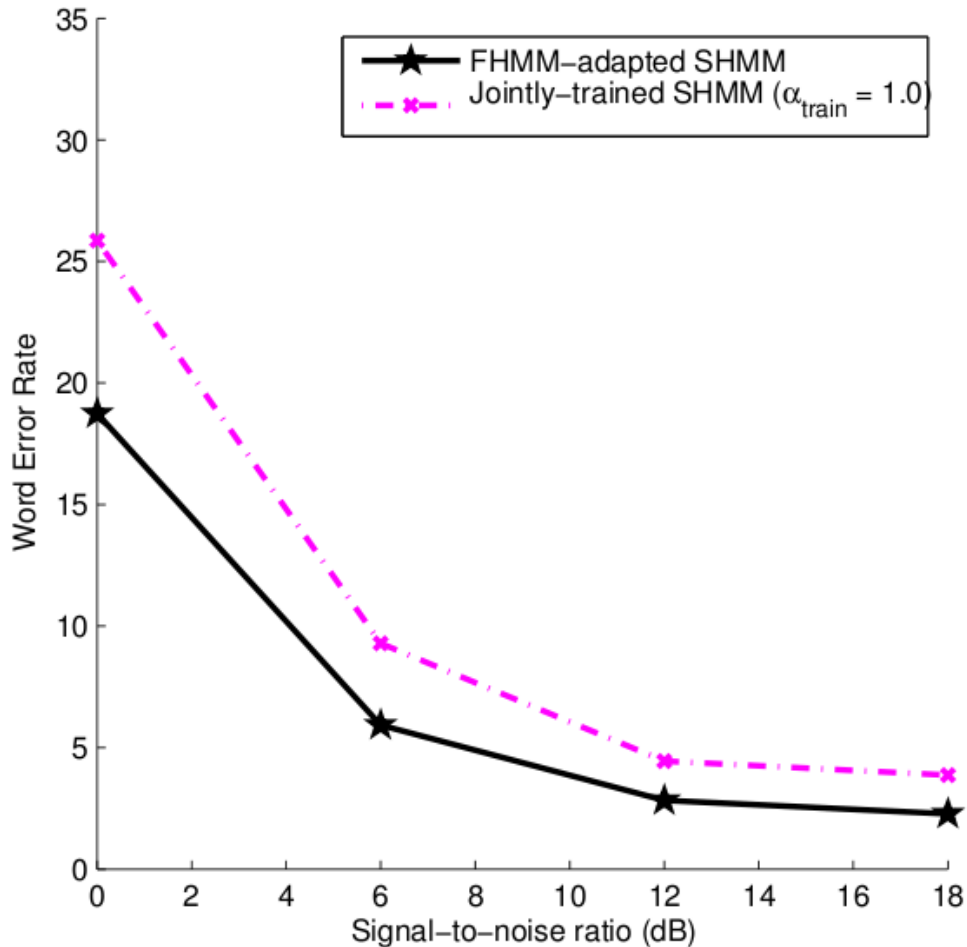
# Continuing Research

- We found that the training weight for jointly-trained SHMMs had little effect (Dean et al, 2007)



D. Dean, P. Lucey, S. Sridharan, and T. Wark, "Weighting and normalisation of synchronous HMMs for audio-visual speech recognition," in *Auditory-Visual Speech Processing 2007*, 2007

# Continuing Research



- FHMM-adaptation outperforms jointly-trained, even if the jointly-trained only used audio for alignment during training ( $\alpha_{\text{train}} = 1$ )
- FHMM-adaptation improvement comes about due to initialisation errors in joint-training

# Possible Avenues of Future Research

- Train AHMMs using a similar technique
  - determine tie-points using audio alone
- Adapting SHMMs from external audio models
  - While there are many, and large, audio speech databases available, audio-visual databases are thinner on the ground
  - FHMM-adaptation could be used to adapt an audio-visual SHMM from an external audio HMM

# Questions?

