Papers that I found on the web

“Event-Driven Deep Neural Network Hardware System for Sensor Fusion”
Comme nous, il utilise une capacite computationnelle onboard limité. Il fait un DNN pour fusionner l'audio et le video pour reconnaitre des chiffres.
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7539099

"Multimodal Deep Learning"
Fusionne audio et video pour améliorer les performance dans un DNN. Je le vois souvent cité.
http://ai.stanford.edu/~ang/papers/icml11-MultimodalDeepLearning.pdf

"Multimodal integration learning of robot behavior using deep neural networks"
Fusionne audio et video, mais avec un robot physique avec des capacités motrices.
http://www.sciencedirect.com/science/article/pii/S0921889014000396

"Towards a Multimodal Sensorimotor Coordination Based Object Recognition System"
Fusionne video, courant de moteur et pression. Robot très mobile (quatre pattes). Met l'accent sur la manipulation pour apprendre un objet.
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4141927

“Multimodal Gesture Recognition using Multi-stream Recurrent neural Network”
Authors use and MRNN with LSTM-RRNs to consider temporal dynamics from their multimodal dataset. They out-perform state of the art methods in the SKIG dataset with 97% accuracy. Very recent paper
http://www.nlab.ci.i.u-tokyo.ac.jp/pdf/psivt2015.pdf

“Audio/Video Fusion for objects recognition”
Authors use speciallized techniques for image processing and audio processing and fuse them after wards. They test their method with moving toys. They focus on the recognition of moving objects and their associated sounds.
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5354442

“Multimodal object Categorization by a Robot”
Uses audio, camera, and robotic arm/hand, with haptic feedback for multimodal input to categorize common objects (spoon, marracas, teddy bear. Etc) in a cluttered environnement. Categorisation method is pLSA. Overall, haptic input helped over only visual.
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4399634

“Multimodal Dimensional Affect Recognition using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks”
Uses a LSTM which is bidirectional and uses a moving average filter at different step to smooth out output spikes (DBLSTM-MA). Used on the AVEC2014 database to recognize emotion from audio/video samples. They score higher or comparable to other state-of-the-art methods.
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7344573

Provided by Mr. Brodeur

“A scalable unsupervised Deep Multimodal Learning System”
Multimodal challenges:
- Differences in signal complexity of the channels
- Scaling the training time and memory with respect to number of channels

Supervised tehcniques dont scale well with multiple modalities because they require fine tuning
4 modalities : audio, images, class and motor
Problems:
- Training all modality combinations requires too much time
- One channel will dominate all others in terms of the associative memory

(Associative memory = top layer RBM that covers all DBNs)
Solution in wake-sleep algorithm which fine tunes RBM weights in linear time
Differences induced:
- 2 sets of weights per connection. One for each direction
- <vi hj >^tc calculated during the top-down pass until error is satisfactory

Benefits:
1- Error associated with generative weights of one channel is factored
2- Algorithm is linear and scales well

Results
Reconstruct Class : from image-> 4.1% err from audio -> 14% err
Reconstruct Image : frim audio -> 19%
Reconstruct Audio

Input\ Reconstucts	Class	Image	Audio	Motor
Class	0	0	0	1.43% err
Image	4.1% err	N/A	19% err	1.63% err
Audio	14% err	19% err	N/A	1.92 % err
Motor	0	0	11% err	N/A

Result notes:
authors noted '4' and '5' are spoken longer than '6' or '8' so the STFT had a negative effect on them.
Im not sure how error was calculated for motor reconstruction <2% errors seems very generous for really crude letters reconstructions. See figure 7 for what I mean.
http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS16/paper/download/12928/12540