PREDICTION-BASED AUDIOVISUAL FUSION

Stavros Petridis

London 12/1/2015

Audiovisual Fusion

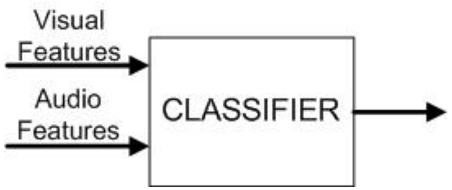
- Goal: Combine information carried by audio and visual modalities.

- In most applications the audio modality is the most informative. The video modality contains information which is:
  - Redundant
  - Complementary

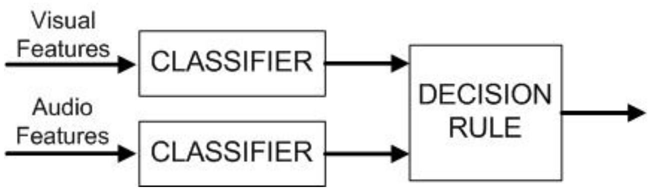
- Research in:
  - Psychology
  - Neuroscience
  - Computer Science

Types of Fusion

- Feature Level



- Decision Level

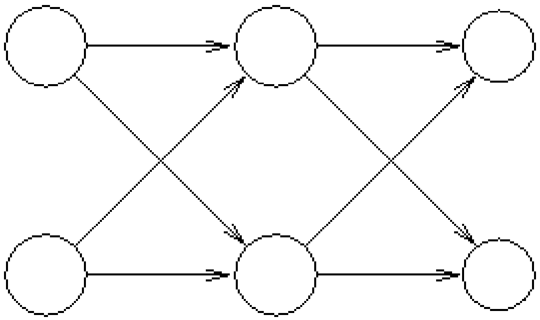


- Model/Classifier/Mid-Level

  e.g., Coupled HMMs,

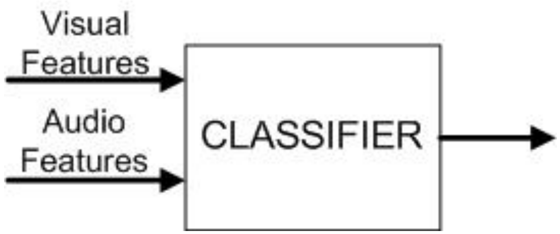
  Multistream HMMs

  Multistream Fused HMMs



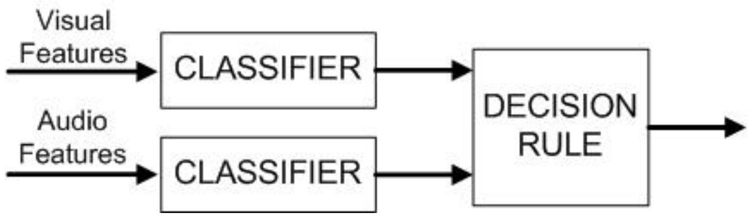
(c) Coupled HMM

Feature-Level Fusion



- Takes into account the spatiotemporal relationship between the audio and visual features, i.e., it models the co-evolution of the audio/visual features
- Requires synchronisation (usually audio/visual features are extracted at different frame rates)
- Increases the dimensionality
- After training the relative weights of each stream cannot change as they are determined internally by the classifier.

Decision-Level Fusion

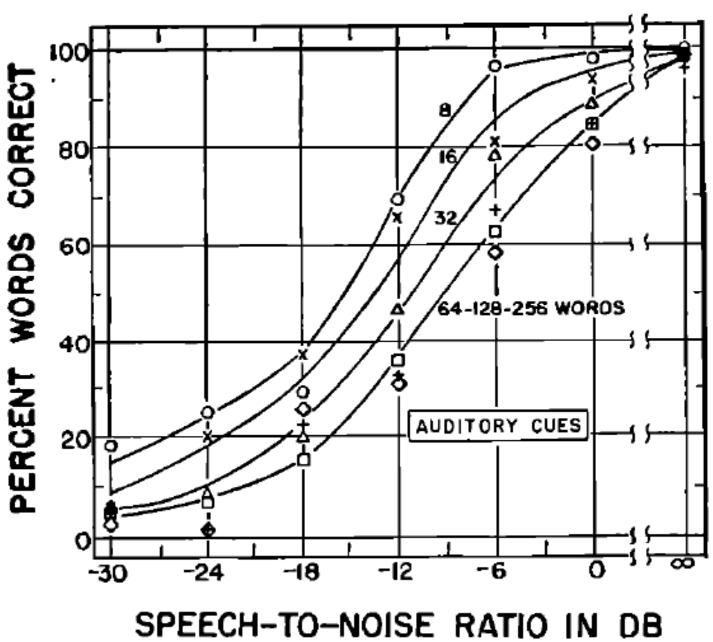


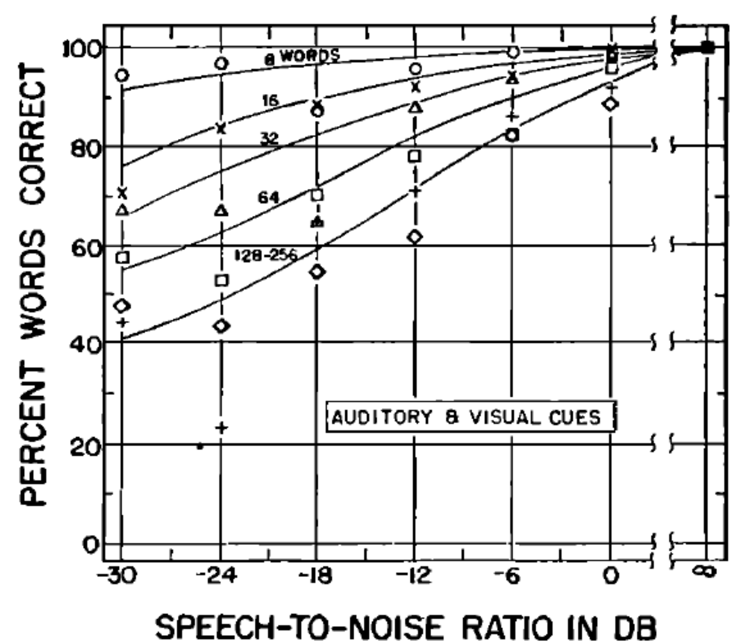
- Modalities are processed independently
- Requires training of multiple classifiers
- Does not require synchronisation
- Dimensionality does not increase
- Relative weights of each stream can easily change by adjusting the weights.

Imperial College
London

Research in Psychology

- Speech becomes more audible when facial movements are visible
  - Visual signal -> 6 – 18 dB gain in SNR
    [W.H. Sumby, I. Pollack (1954), Visual contribution to speech intelligibility in noise,]

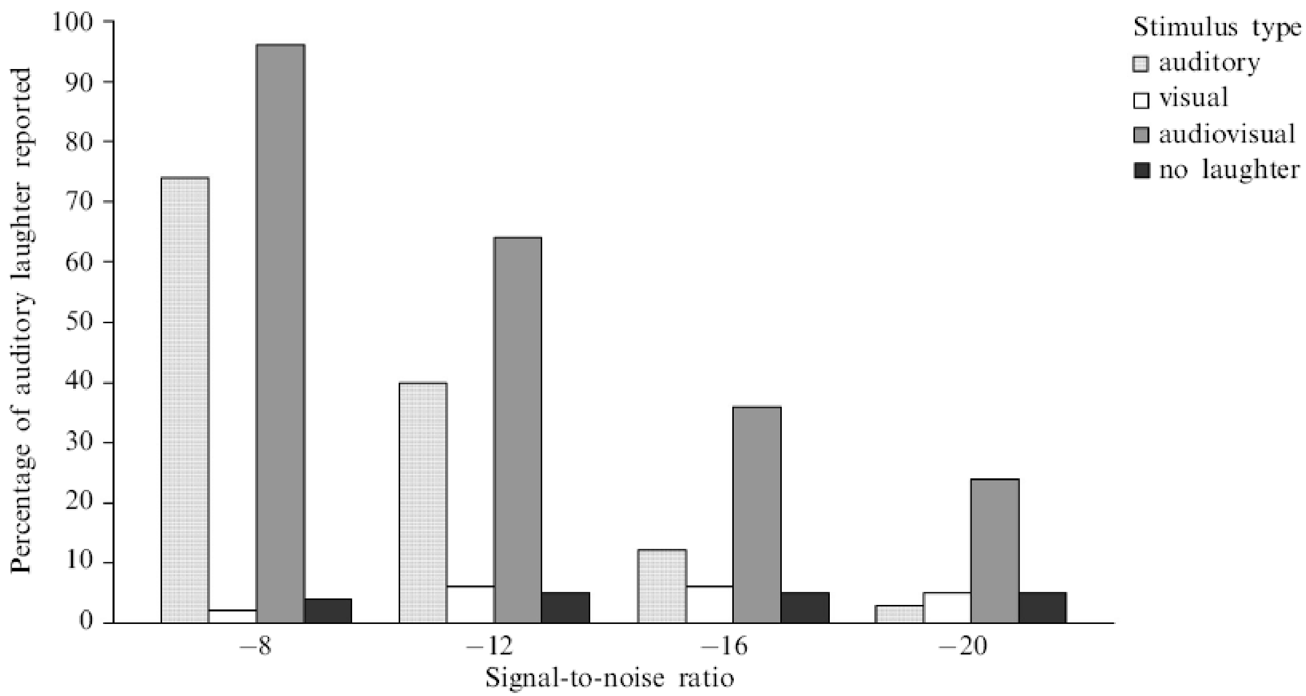




Research in Psychology

- Laughter becomes more audible when facial movements are visible

[T. R. Jordan, L. Abedipour, (2010), The importance of laughing in your face: Influences of visual laughter on auditory laughter perception]



Imperial College London

Research in Psychology

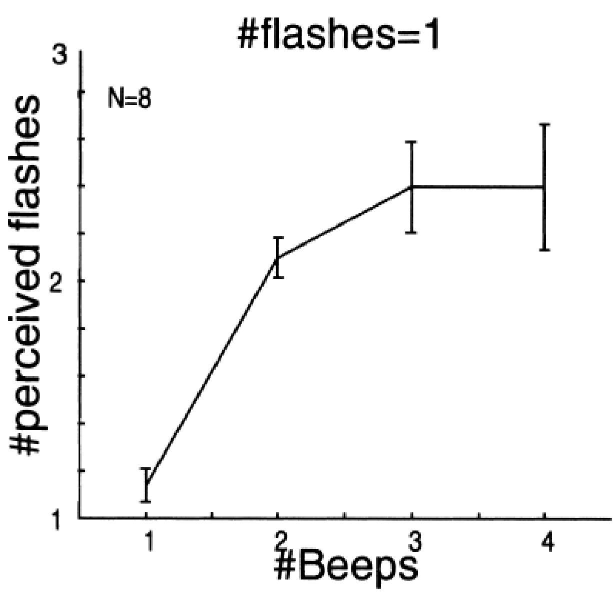
- McGurk Effect
  - The auditory component of one sound is paired with the visual component of another sound, leading to the perception of a third sound
  - Interaction between vision and hearing
  - Vision can alter the perception of sounds

[McGurk, H & MacDonald, J (1976); Hearing lips and seeing voices]

Research in Psychology

- Sound-induced flash illusion
  - Hearing can alter visual perception

[L. Shams, Y. Kamitani, S. Shimojo (2002); Visual illusion induced by sound]



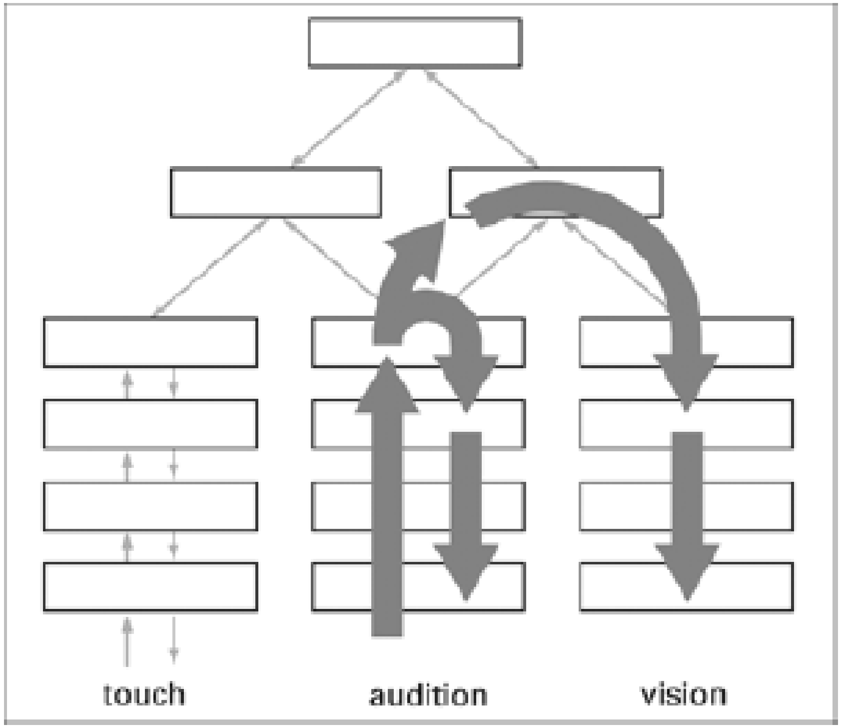
http://www.cns.atr.jp/~kmtn/soundInducedIllusoryFlash2/

Imperial College London

Prediction-based Fusion - Motivation

- Memory-Prediction Framework [J. Hawkins (2004), On Intelligence]

  - Predict what we will hear / see based on what we see / hear



Imperial College London

Prediction-based Fusion - Motivation

- Relationship between acoustic and visual features (speech)
  - A->V mapping: correlation 0.7 – 0.85

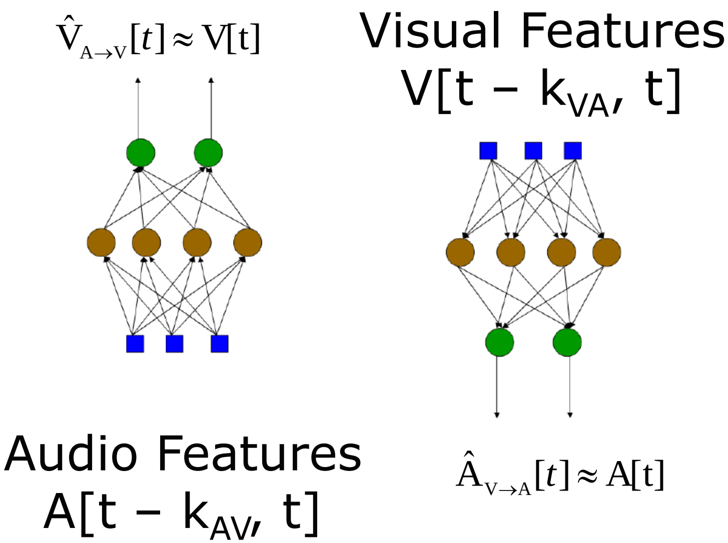
- Reasonable to assume that:

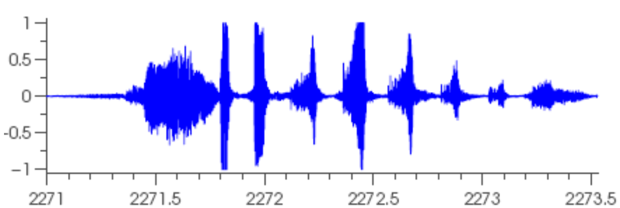
  1) Relationship between audio and visual features is different in speech and laughter (or other non-linguistic vocalisations)
  2) Time evolution of audio and visual features is different in speech and laughter (or other non-linguistic vocalisations)

- We can learn the AV relationship (i.e., learn the mapping between A and V) for each class. Classify an example based on which mapping better describes a new example.

Imperial College London

Prediction-Based Fusion – Cross Prediction Component





- For each class c learn the mapping *f* between audio and visual features

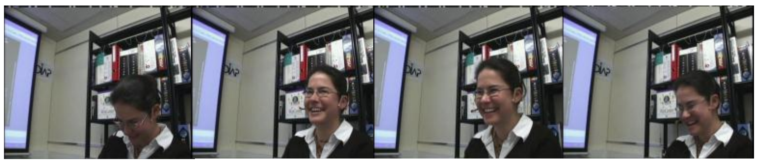
$$f^c_{A\rightarrow V}(A^c[t-k^c_{AV}, t]) = \hat{V}^c_{A\rightarrow V}[t] \approx V^c[t]$$

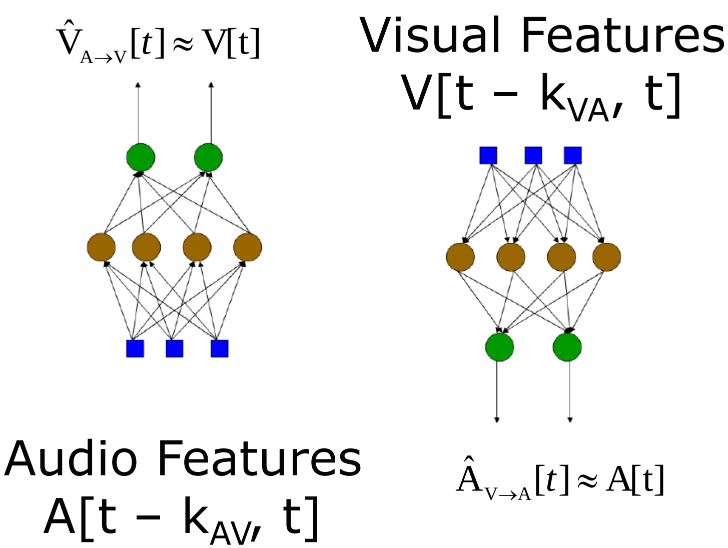
$$f^c_{V\rightarrow A}(V^c[t-k^c_{VA}, t]) = \hat{A}^c_{V\rightarrow A}[t] \approx A^c[t]$$

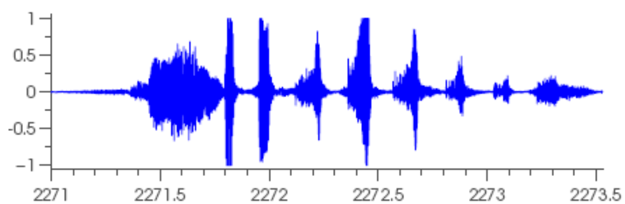
- This corresponds to feature-level fusion where concatenation is replaced by the AV mapping functions

Imperial College London

Prediction-Based Fusion – Cross Prediction Component







- Classification: The audio/visual features are fed to the AV mapping functions already learned (one set of functions for each class)

$$f^c_{A\to V}(A^c[t - k^c_{AV}, t]) = \hat{V}^c_{A\to V}[t]$$

$$f^c_{V\to A}(V^c[t - k^c_{VA}, t]) = \hat{A}^c_{V\to A}[t]$$

- The prediction error over the entire sequence is computed.

$$e^c_{A\to V} = \sum_{i=1}^{N} Err(\hat{V}^c_{A\to V}[i], V[i])$$

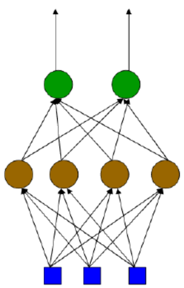
$$e^c_{V\to A} = \sum_{i=1}^{N} Err(\hat{A}^c_{V\to A}[i], A[i])$$

- Error: MSE, MAE, L2

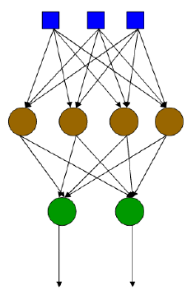
Imperial College London

Prediction-Based Fusion

$\hat{V}_{A \to V}[t] \approx V[t]$

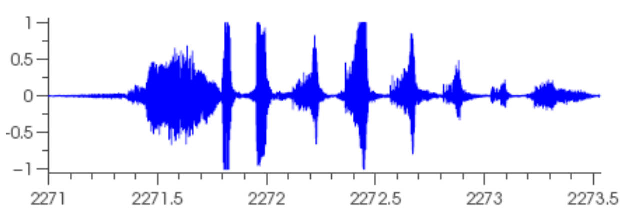


Visual Features
$V[t - k_{VA}, t]$

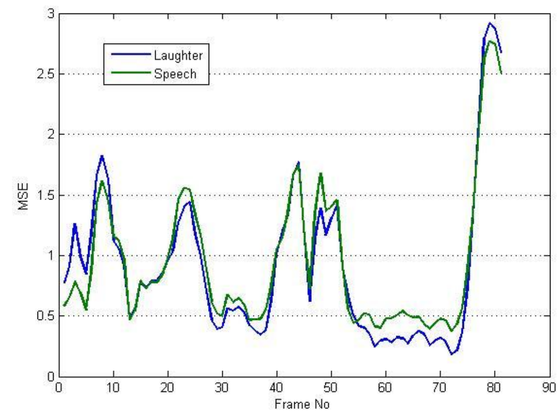


Audio Features
$A[t - k_{AV}, t]$

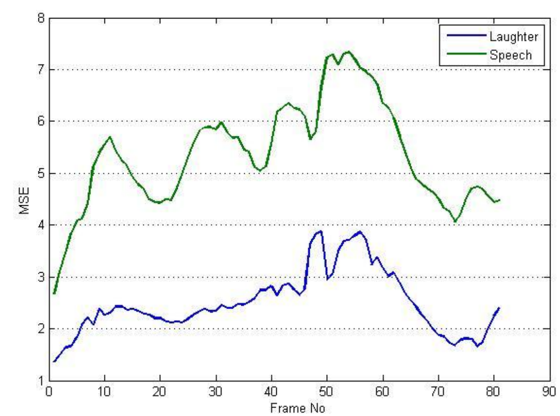
$\hat{A}_{V \to A}[t] \approx A[t]$



Video-to-Audio Mapping

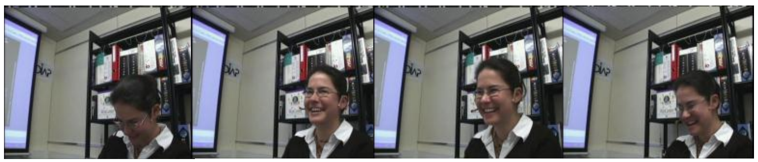


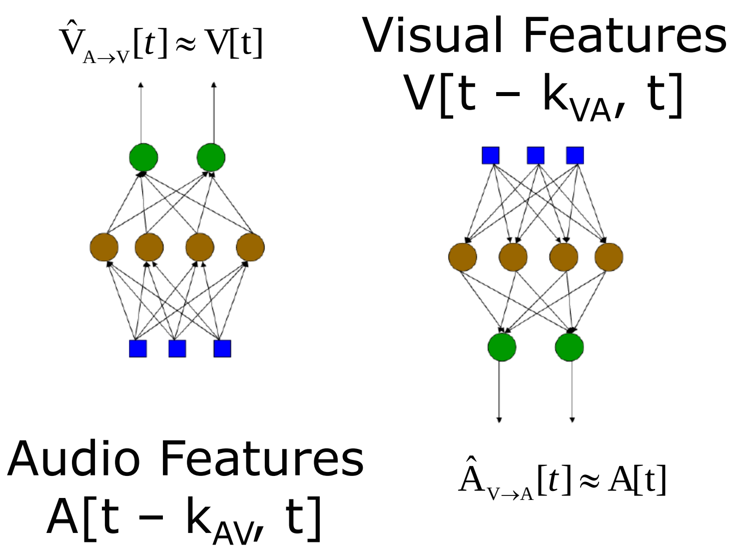
Audio-to-Video Mapping

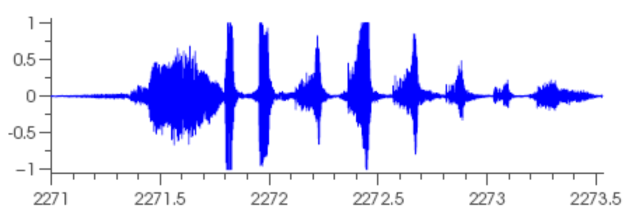


Imperial College London

Prediction-Based Fusion – Cross Prediction Component







- The prediction errors for each class can be combined

$$e^c_{CP} = w^c_{AV} \times e^c_{A\to V} + w^c_{VA} \times e^c_{V\to A}$$

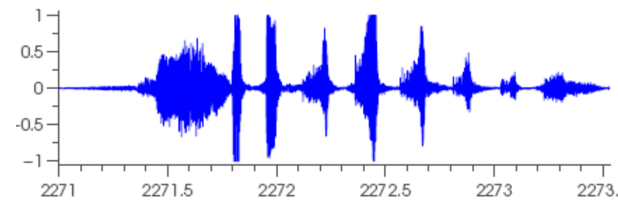
$$w^c_{AV} + w^c_{VA} = 1$$

- The sequence is labelled based on the predictor which corresponds to the lowest prediction error, i.e., class-specific predictor that best explains the AV relationship.

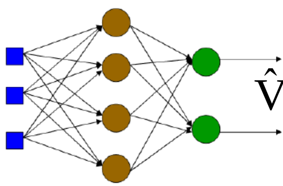
$$PredictedClass = \arg\min_{c=1\ldots C} e^c$$

- The main idea is that the predictors which have been trained on the correct class will produce a lower prediction error .

Prediction-Based Fusion – Intra-Prediction Component

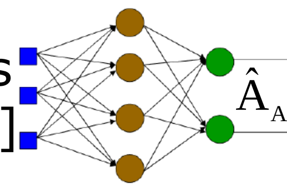


Visual Features
$V[t - k_{VV}, t - 1]$



$\hat{V}_{V\to V}[t] \approx V[t]$

Audio Features
$A[t - k_{AA}, t - 1]$



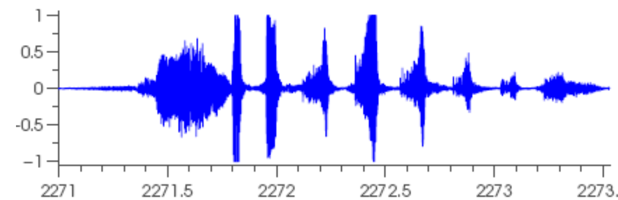
$\hat{A}_{A\to A}[t] \approx A[t]$

- For each class c learn the mapping *f* between past audio / visual and future audio / visual features.
- This corresponds to decision-level fusion.

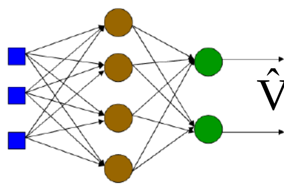
$$f^c_{A\to A}(A^c[t-k^c_{AA}, t-1]) = \hat{A}^c_{A\to A}[t] \approx A^c[t]$$

$$f^c_{V\to V}(V^c[t-k^c_{VV}, t-1]) = \hat{V}^c_{V\to V}[t] \approx V^c[t]$$

Prediction-Based Fusion – Intra-Prediction Component

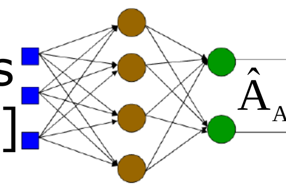


Visual Features
$V[t - k_{VV}, t - 1]$



$\hat{V}_{V\rightarrow V}[t] \approx V[t]$

Audio Features
$A[t - k_{AA}, t - 1]$



$\hat{A}_{A\rightarrow A}[t] \approx A[t]$

- Classification: The audio/visual features are fed to the AV mapping functions already learned (one set of functions for each class)

$$f^c_{A\rightarrow A}(A^c[t - k^c_{AA}, t - 1]) = \hat{A}^c_{A\rightarrow A}[t]$$

$$f^c_{V\rightarrow V}(V^c[t - k^c_{VV}, t - 1]) = \hat{V}^c_{V\rightarrow V}[t]$$

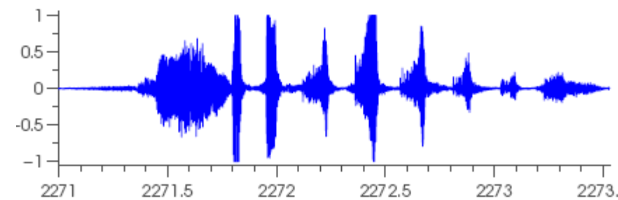
- The prediction error over the entire sequence is computed.

$$e^c_{A\rightarrow A} = \sum_{i=1}^{N} Err(\hat{A}^c_{A\rightarrow A}[i], A[i])$$

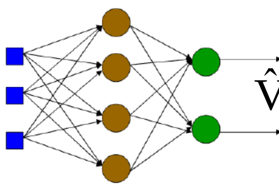
$$e^c_{V\rightarrow V} = \sum_{i=1}^{N} Err(\hat{V}^c_{V\rightarrow V}[i], V[i])$$

Imperial College London

Prediction-Based Fusion – Intra-Prediction Component

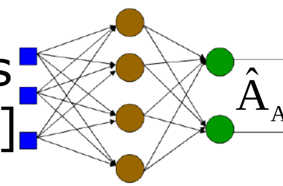


Visual Features
$V[t - k_{VV}, t - 1]$



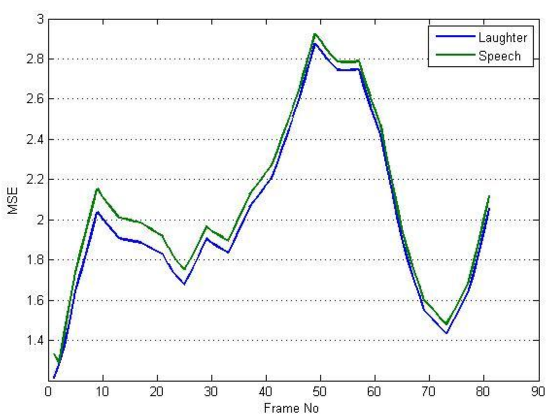
$\hat{V}_{V \to V}[t] \approx V[t]$

Audio Features
$A[t - k_{AA}, t - 1]$

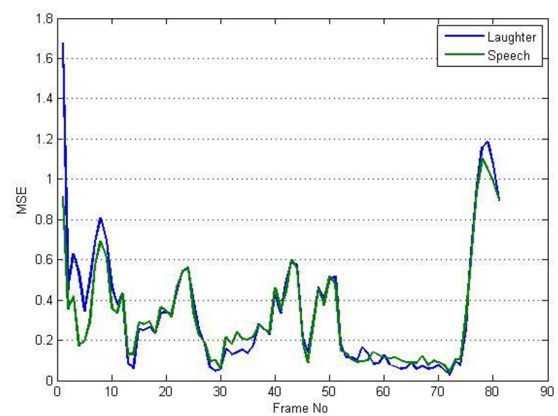


$\hat{A}_{A \to A}[t] \approx A[t]$

Video-to-Video Mapping

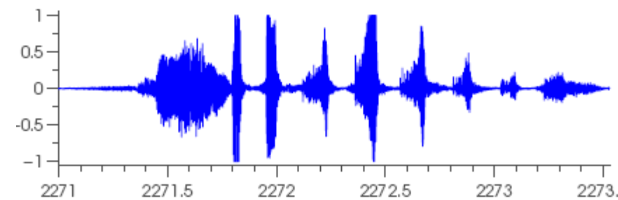


Audio-to-Audio Mapping

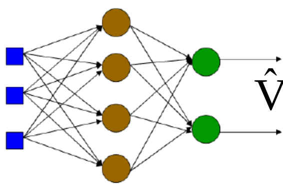


Imperial College London

Prediction-Based Fusion – Intra-Prediction Component

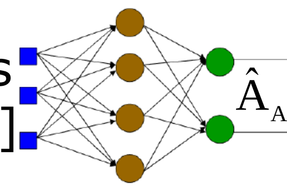


Visual Features
V[t – k_{VV}, t - 1]



$\hat{V}_{V\to V}[t] \approx V[t]$

Audio Features
A[t – k_{AA}, t - 1]



$\hat{A}_{A\to A}[t] \approx A[t]$

- The prediction errors for each class can be combined

$$e^c_{IP} = w^c_{AA} \times e^c_{A\to A} + w^c_{VV} \times e^c_{V\to V}$$

$$w^c_{AA} + w^c_{VV} = 1$$

- The sequence is labelled based on the predictor which corresponds to the lowest prediction error, i.e., class-specific predictor that best explains the AV relationship.

Prediction-Based Fusion – Final System

- The cross-prediction and intra-prediction modules can also be combined

$$e^c = w^c_{CP} \times e^c_{CP} + w^c_{IP} \times e^c_{IP}$$

$$w^c_{CP} + w^c_{IP} = 1$$

- The sequence is labelled based on the predictor which corresponds to the lowest prediction error, i.e., class-specific predictor that best explains the AV relationship.

$$PredictedClass = \arg\min_{c=1...C} e^c$$

- The main idea is that the predictors which have been trained on the correct class will produce a lower prediction error .

Prediction-based Fusion

$$e^c = w^c_{CP} \times e^c_{CP} + w^c_{IP} \times e^c_{IP}$$

$$w^c_{CP} + w^c_{IP} = 1$$

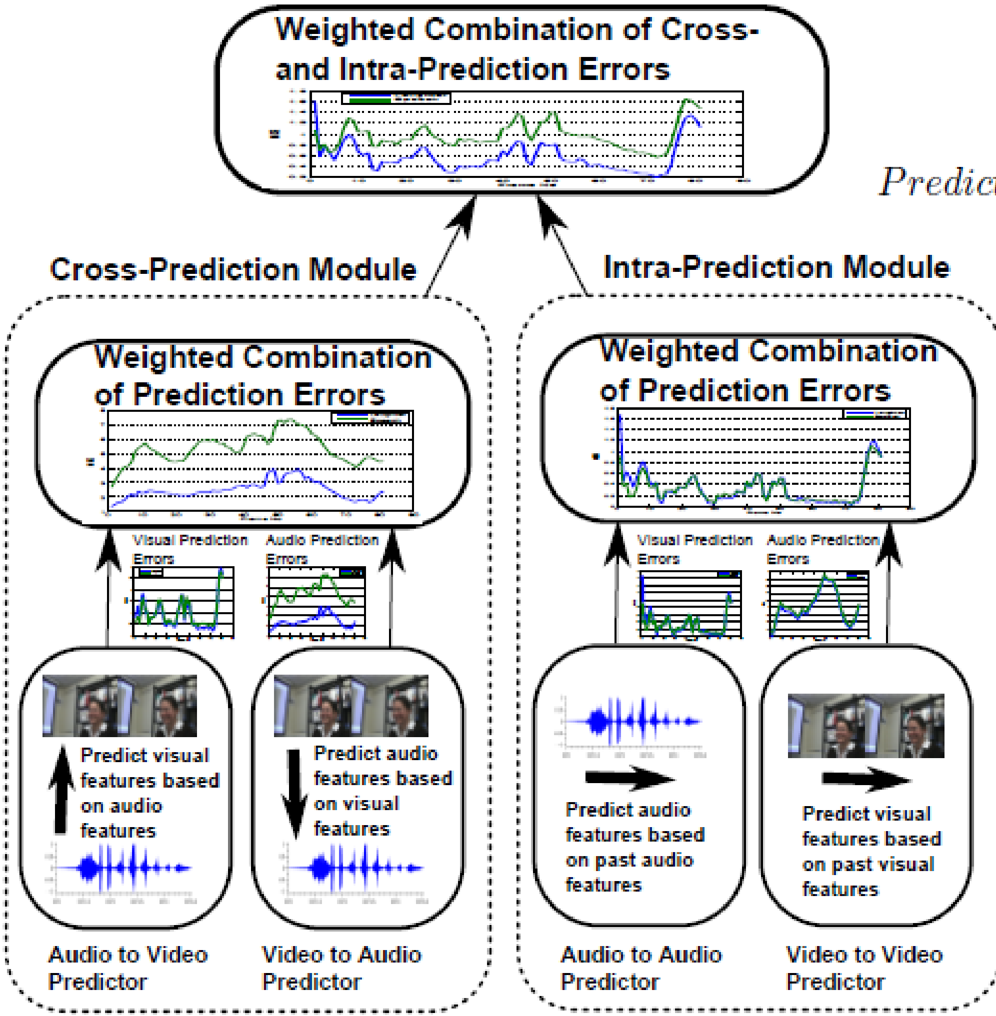
$$PredictedClass = \arg\min_{c=1\ldots C} e^c$$

$$e^c_{CP} = w^c_{AV} \times e^c_{A\to V} + w^c_{VA} \times e^c_{V\to A}$$

$$w^c_{AV} + w^c_{VA} = 1$$

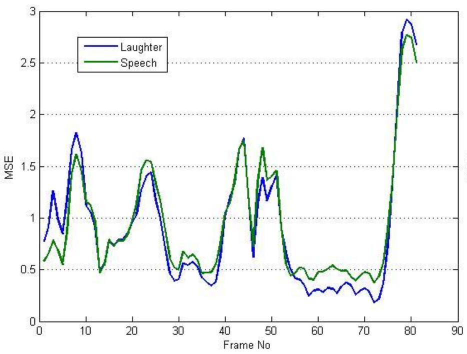
$$e^c_{IP} = w^c_{AA} \times e^c_{A\to A} + w^c_{VV} \times e^c_{V\to V}$$

$$w^c_{AA} + w^c_{VV} = 1$$

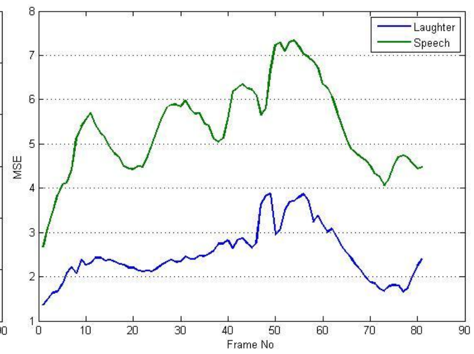


Imperial College London

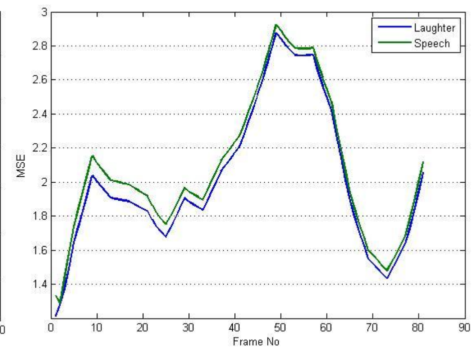
Weights Normalisation



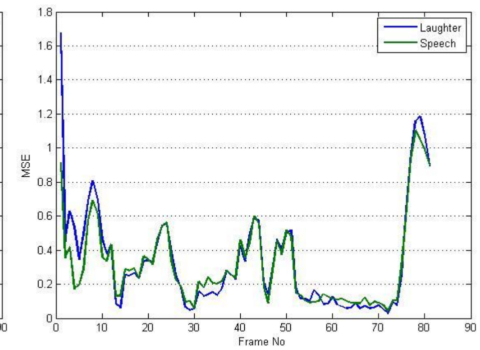
V2A



A2V



V2V



A2A

- Errors are in different scale.
- Weights do not reflect only the relative importance but also take into account scaling differences.
- Errors can be normalised, e.g. softmax

Datasets

- AMI, SAL, MAHNOB: Laughter/Speech
- AVIC: Laughter, Hesitation, Consent, Garbage

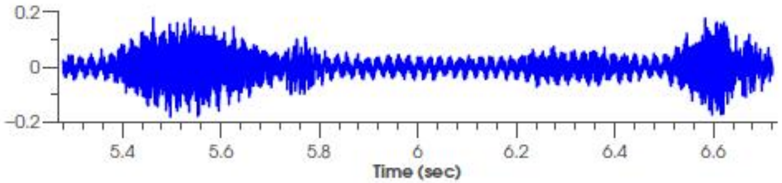
- Cross-database experiments for laughter/speech
  - Train: SAL (10 subjects)
  - Val: SAL (5 subjects)
  - Test: MAHNOB

- AVIC is divided into training/validation/test sets (8 subj. each)

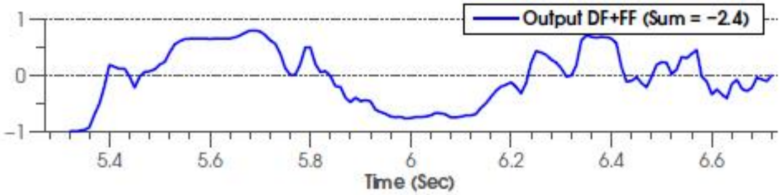
- Visual features: PCA on points
- Audio features: MFCCs

Example

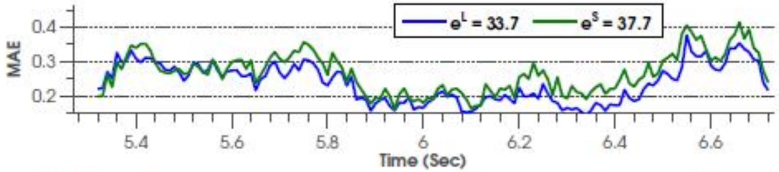
(a) Frame 133 (b) Frame 145 (c) Frame 157 (d) Frame 168



(e) Audio signal



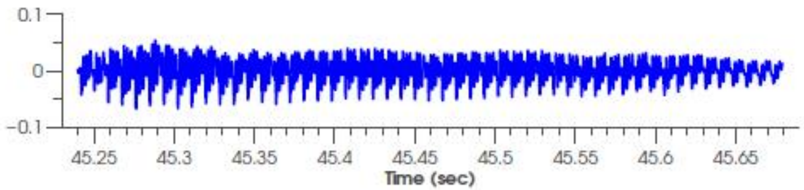
(f) Output of DF + FF. The caption shows the total score. The example is misclassified as speech since the total score is negative.



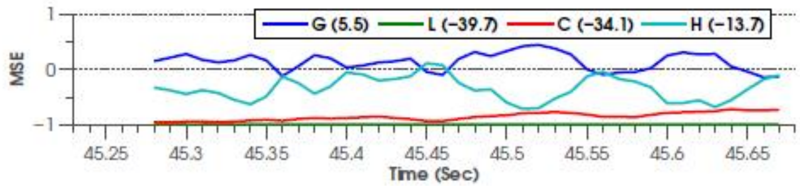
(g) MAE of the laughter and speech models. The caption shows the total MAE over the entire episode. The example is classified as laughter since this model leads to the lowest error.

Example

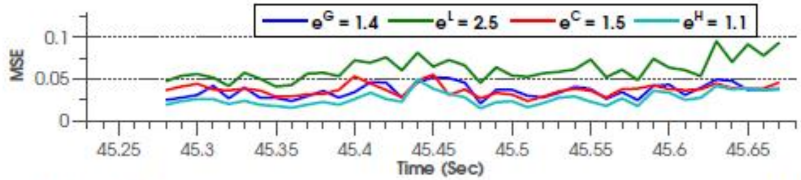
(a) Frame 929 (b) Frame 934 (c) Frame 940 (d) Frame 945



(e) Audio signal



(f) Output of DF + FF. The caption shows the total score. The example is misclassified as speech since the speech output leads to the highest score.



(g) MSE of the laughter and speech models. The caption shows the total MSE over the entire episode. The example is classified as laughter since this model leads to the lowest error.

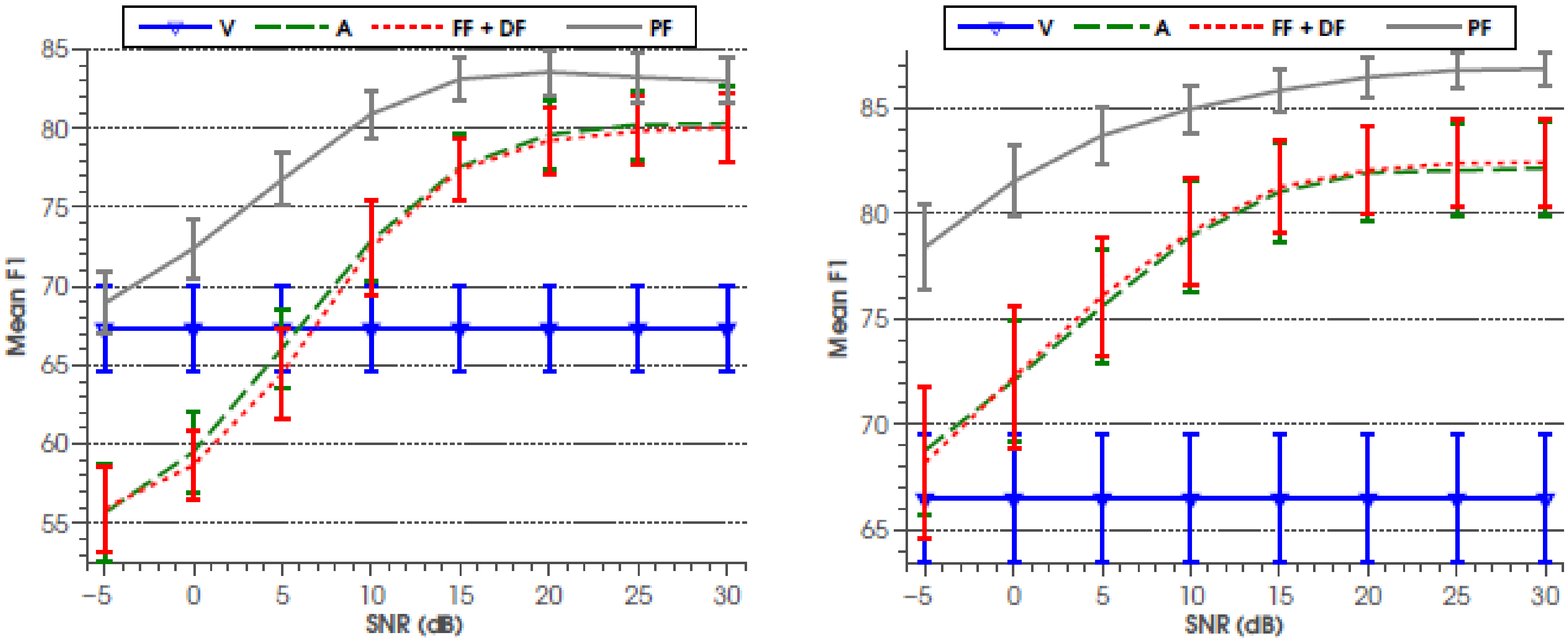
Results

Classification System Test →	F1 Laughter	F1 Speech	F1 Mean	CR	UAR	F1 Laughter	F1 Speech	F1 Mean	CR	UAR
	AMI					MAHNOB				
A	73.7 (3.4)	85.3 (1.4)	79.5 (2.4)	81.1 (2.0)	79.0 (2.2)	76.2 (3.3)	88.2 (1.1)	82.2 (2.2)	84.2 (1.7)	80.8 (2.2)
V	58.5 (5.2)	76.1 (1.0)	67.3 (2.8)	69.8 (1.7)	67.7 (2.2)	55.0 (5.6)	78.0 (1.0)	66.5 (3.0)	70.5 (1.8)	66.3 (2.9)
A + V (PF - S)	76.6 (1.9)†	86.2 (0.7)†	81.4 (1.3)†	82.6 (1.1)†	80.8 (1.2)†	**83.5 (1.2)**†	**90.4 (0.5)**†	**86.9 (0.8)**†	**87.8 (0.7)**†	**86.0 (1.0)**†
A + V (PF - N)	**79.4 (2.2)**†	**87.6 (1.0)**†	**83.5 (1.6)**†	**84.5 (1.4)**†	**82.9 (1.5)**†	**84.7 (2.2)**†	**91.1 (0.9)**†	**87.9 (1.6)**†	**88.7 (1.3)**†	**87.0 (1.7)**†
A + V (DF + FF)	73.5 (2.9)	85.4 (1.1)	79.5 (2.0)	81.2 (1.7)	79.0 (1.8)	76.5 (3.2)	88.4 (1.1)	82.5 (2.1)	84.5 (1.7)	81.0 (2.1)

Classification System Test →	F1 Garbage	F1 Laughter	F1 Consent	F1 Hesitation	F1 Mean	CR	UAR
	AVIC						
A	51.1 (3.8)	58.3 (2.6)	40.0 (5.2)	67.2 (2.8)	54.1 (2.2)	58.8 (2.4)	58.7 (2.4)
V	44.4 (4.1)	38.9 (2.6)	35.5 (3.4)	57.1 (3.7)	44.0 (2.0)	48.5 (2.6)	48.9 (2.5)
A + V (PF - S)	**56.9 (2.9)**†	71.0 (2.6)†	44.0 (3.3)	75.9 (1.8)†	62.0 (1.6)†	67.7 (1.8)†	64.0 (1.9)
A + V (PF - N)	**57.7 (2.2)**	67.2 (2.5)	**46.2 (4.2)**	74.9 (1.0)	61.5 (1.6)	67.0 (1.2)	**64.2 (2.0)**
A + V (DF + FF)	54.3 (4.0)	60.5 (2.5)	44.8 (5.1)	68.4 (2.7)	57.0 (2.2)	61.1 (2.4)	61.8 (2.2)

Imperial College London

Example



(a) Test Set: AMI (b) Test Set: MAHNOB

High Noise *Low Noise* *Low Noise* *High Noise*

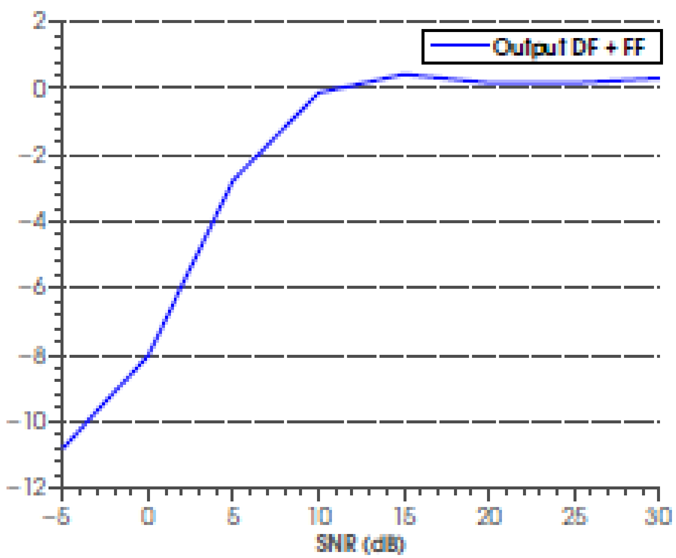
Example

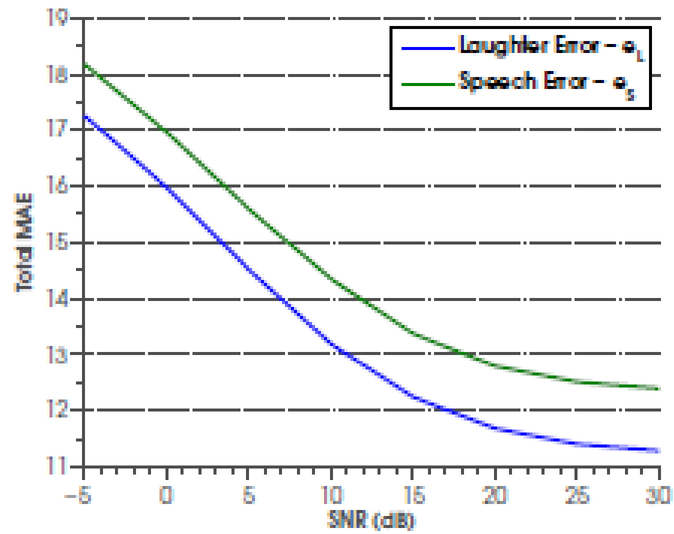
- Laughter example from the MAHNOB DB

- It does not matter if the absolute prediction error increases, what matters is the relative position of the two errors.

Laughter

Speech





High Noise *Low Noise* *High Noise* *Low Noise*

Prediction-based Fusion - Extensions

- Time series clustering
- Segmentation
- Deep NNs

Time Series Clustering

- Cluster examples based on subject
- Train one set of predictors per class for each subject
  - Total No Predictors = NoSubjects x NoClasses
- Label a sequence based on the set of predictors which lead to the lowest prediction error

Table 3.25: Performance measures computed for classification of sequences on the AMI test set using minimum error method.

Performance Measure	Laughter	Speech	Overall
Precision	91.59%	84.80%	88.19%
Recall	79.03%	94.16%	86.59%
F_1 score	84.85%	89.23%	87.04%
Classification Rate	-	-	87.41%

Best on entire
Dataset, mean F1: 80.6

Table 3.27: Performance measures computed for classification of sequences on the MAHNOB test set using minimum error method.

Performance Measure	Laughter	Speech	Overall
Precision	84.14%	88.06%	86.10%
Recall	81.41%	89.94%	85.67%
F_1 score	82.75%	88.99%	85.87%
Classification Rate	-	-	86.56%

Best on entire
Dataset, mean F1: 83.8

Time Series Clustering

- Cluster examples based laughter type, i.e., voiced / unvoiced laughter
- Train one set of predictors per class
- Label a sequence based on the set of predictors which lead to the lowest prediction error. If voiced / unvoiced laughter -> laughter

Table 3.35: Performance measures computed on the AMI test set.

Performance Measure	Laughter	Speech	Overall
Precision	94.34%	86.05%	90.19%
Recall	80.65%	96.10%	88.37%
F_1 score	86.96%	90.80%	88.88%
Classification Rate	-	-	89.21%

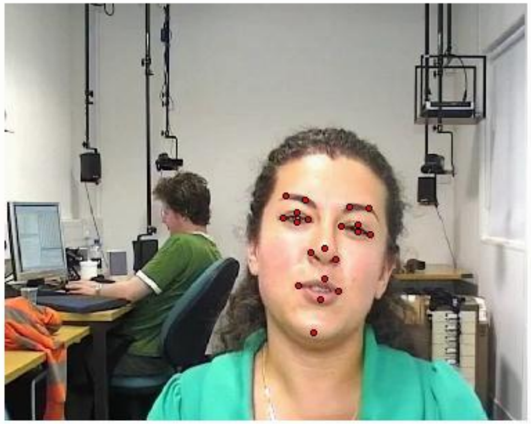
Best on entire Dataset, mean F1: 80.6

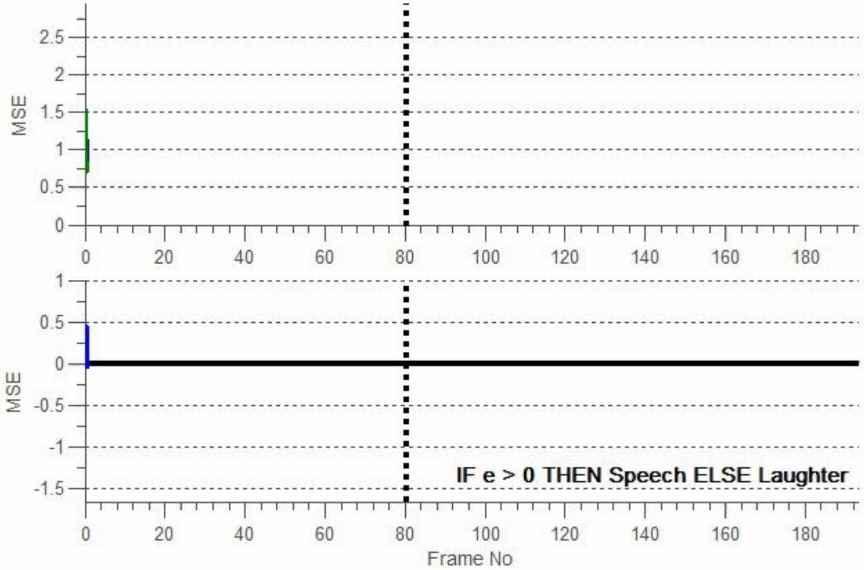
Table 3.37: Performance measures computed on the MAHNOB test set.

Performance Measure	Laughter	Speech	Overall
Precision	91.01%	92.14%	91.57%
Recall	87.73%	94.32%	91.02%
F_1 score	89.34%	93.22%	91.28%
Classification Rate	-	-	91.71%

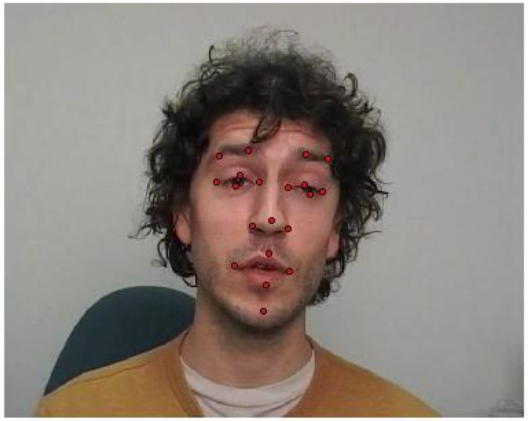
Best on entire Dataset, mean F1: 83.8

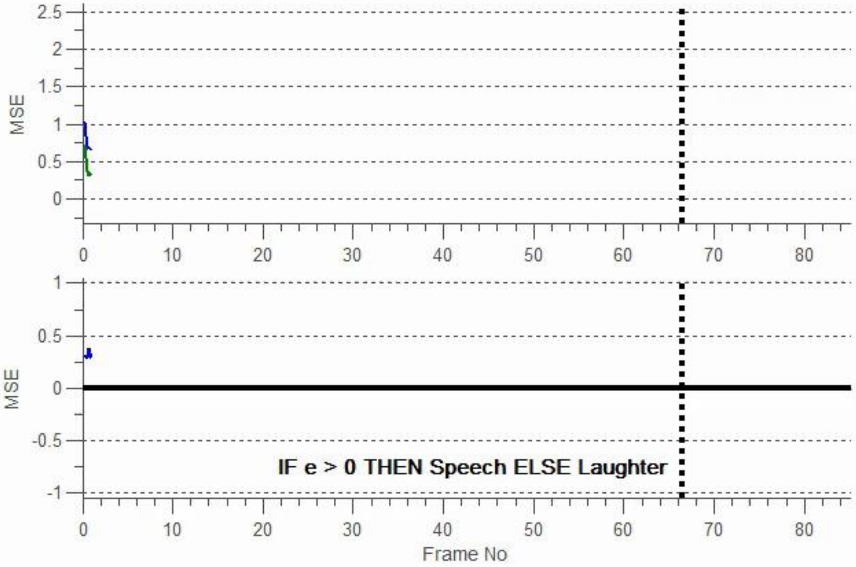
Segmentation – Example 1





Segmentation – Example 2

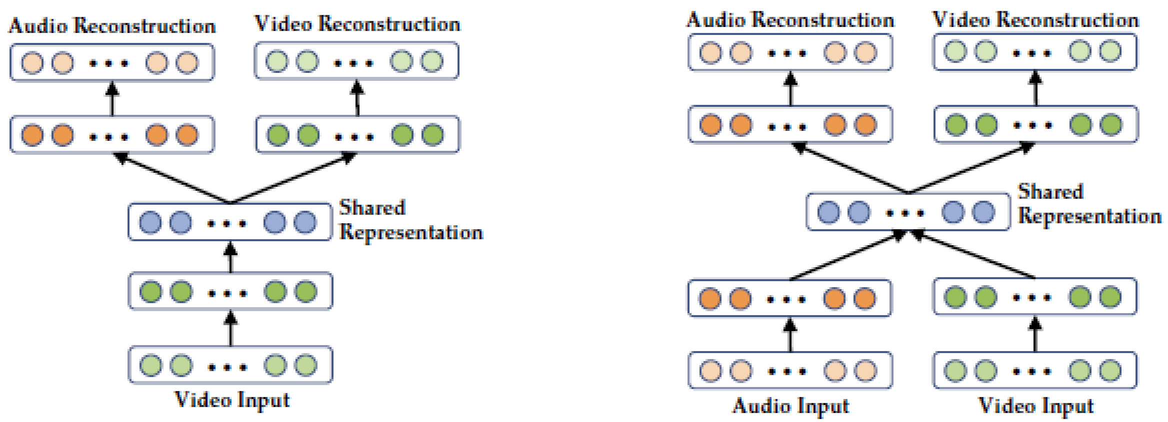




Imperial College London

Prediction-based Fusion - Extensions

- It has been found that visual speech recognition benefits when features are extracted from a deep AE which learns to reconstruct audio features as well.
- Train a DNN to predict Audio Features and future Visual features
- Use bottleneck features for classification, they should model the audiovisual relationship



Ngiam, Jiquan, et al. "Multimodal deep learning." Proceedings of the 28th International Conference on Machine Learning, 2011.

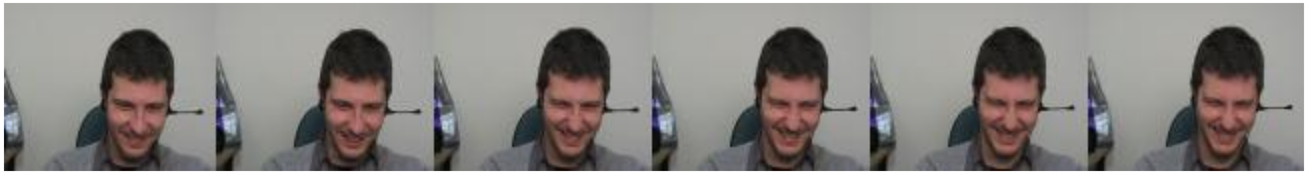
Imperial College London

THANK YOU! ☺

Imperial College London

Datasets

- Elicited Laughter (MAHNOB)



- Dyadic Interaction (AVIC, SAL)

- Meeting Scenario (AMI)

Prediction-based Fusion - Variants

- Comparison of single network-vs-multiple networks
  - Performance is similar

- Comparison of different predictors
  - Prediction-based fusion outperforms DF/FF when NNs, LSTMs, GPs
  - Performance is similar for SVMs, RVMs

- Comparison of different audio feature sets
  - MFCCs, DeltaMFCCs, Pitch, Energy, ZCR
  - Performance is similar