Insight
Centre for Data Analytics

DEEP LEARNING WORKSHOP

Dublin City University
27-28 April 2017

DCU

D2 L9

Audio & Vision

Amaia Salvador
amaia.salvador@upc.edu

PhD Candidate
Universitat Politècnica de Catalunya

UPC

Multimedia

Text

Vision

Audio

Audio & Vision

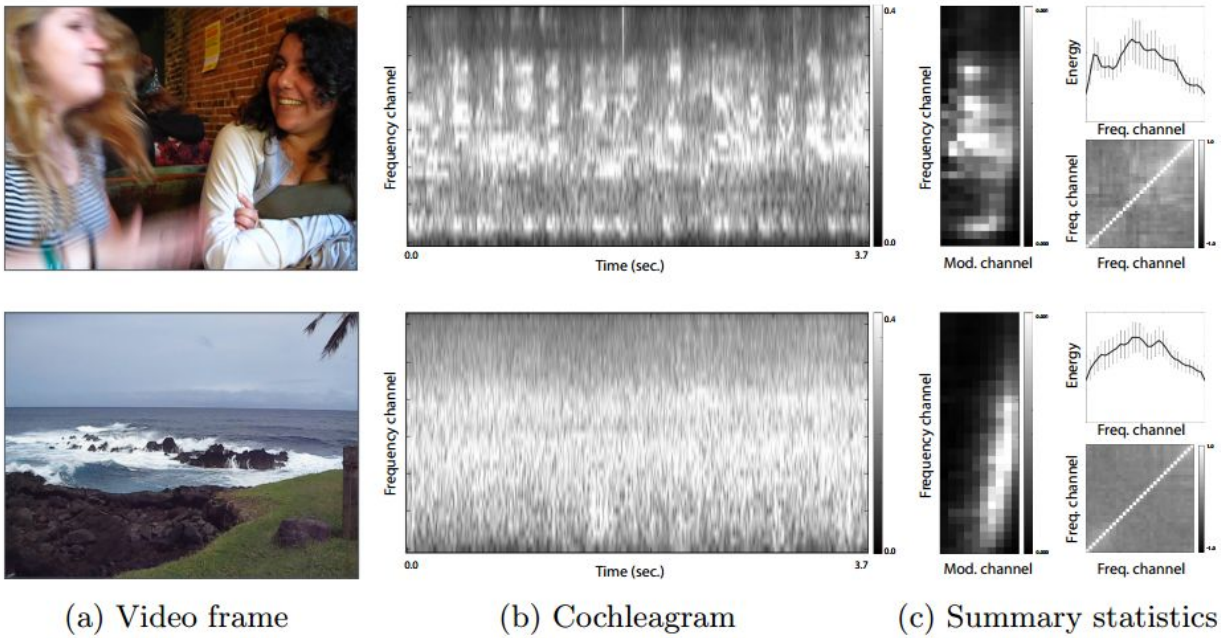
- Transfer Learning
- Sonorization
- Speech generation
- Lip Reading

Audio & Vision

- **Transfer Learning**
- Sonorization
- Speech generation
- Lip Reading

Transfer Learning

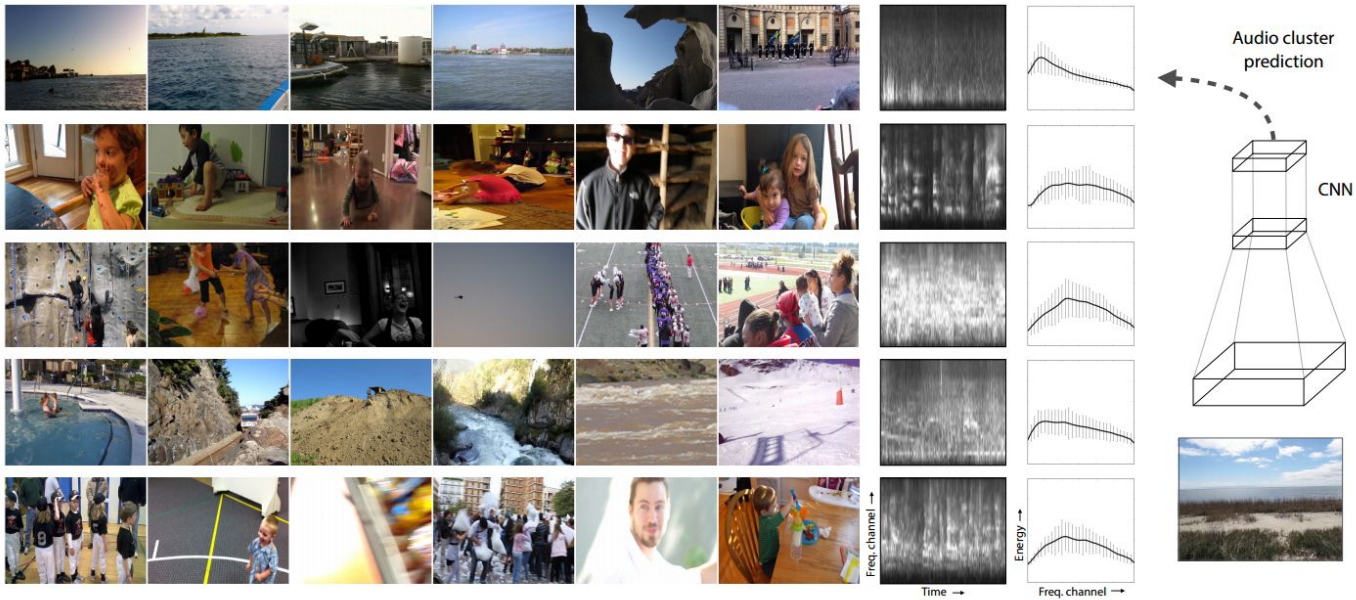
Audio features help learn semantically meaningful visual representation



Owens et al. [Ambient Sound Provides Supervision for Visual Learning.](#) ECCV 2016

Transfer Learning

Images are grouped into audio clusters.
The task is to predict the audio cluster for a given image (classification)

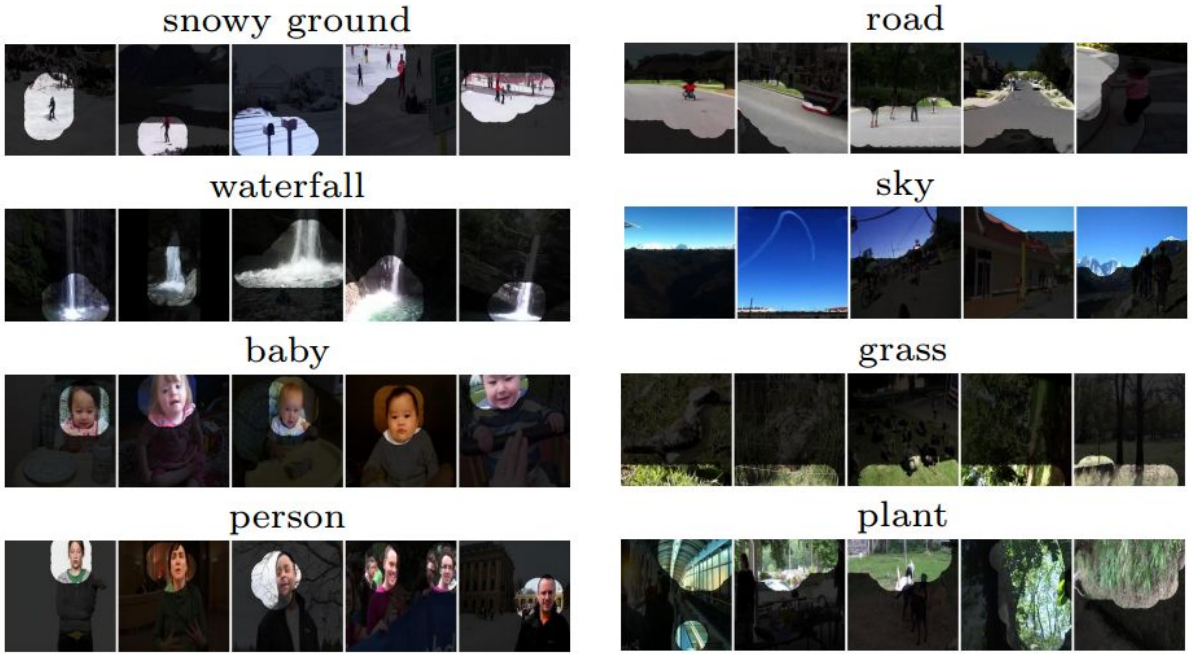


(a) Images grouped by audio cluster (b) Clustered audio stats. (c) CNN model

Owens et al. Ambient Sound Provides Supervision for Visual Learning. ECCV 2016

Transfer Learning

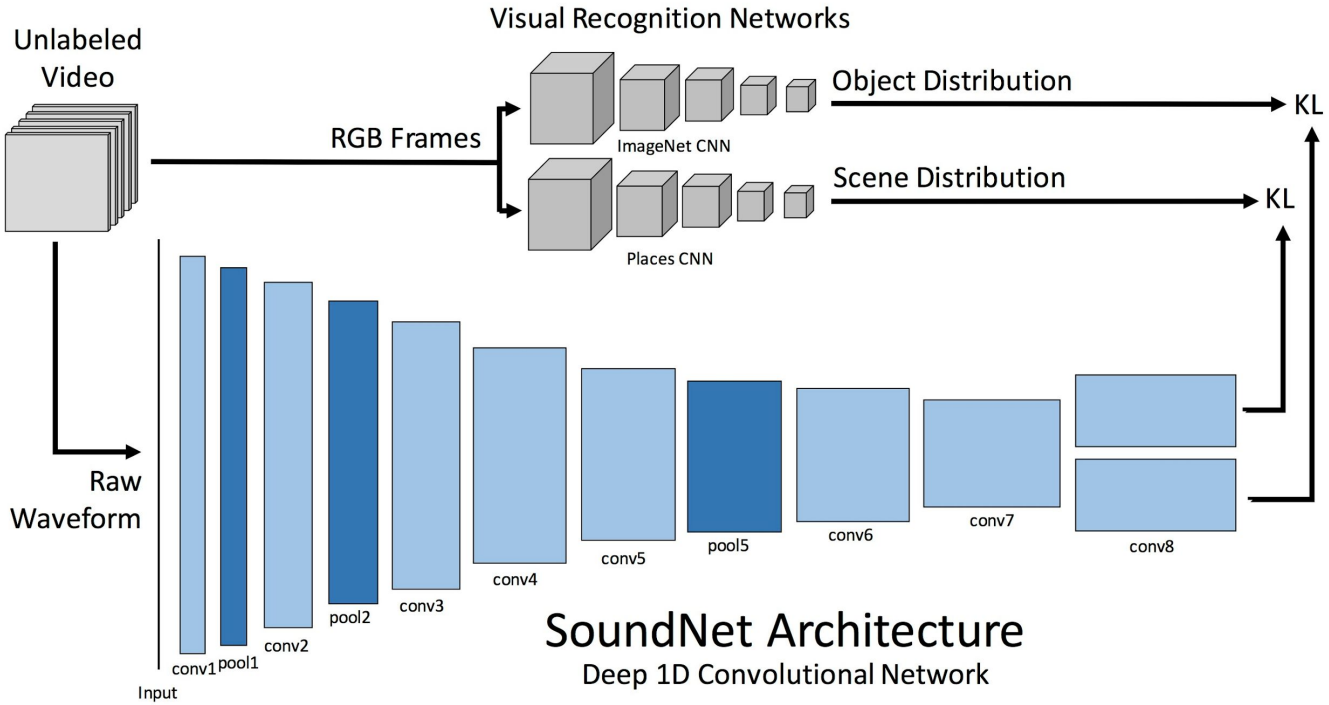
Although the model was not trained with class labels, units with semantical meaning emerge



Owens et al. Ambient Sound Provides Supervision for Visual Learning. ECCV 2016

Transfer Learning: SoundNet

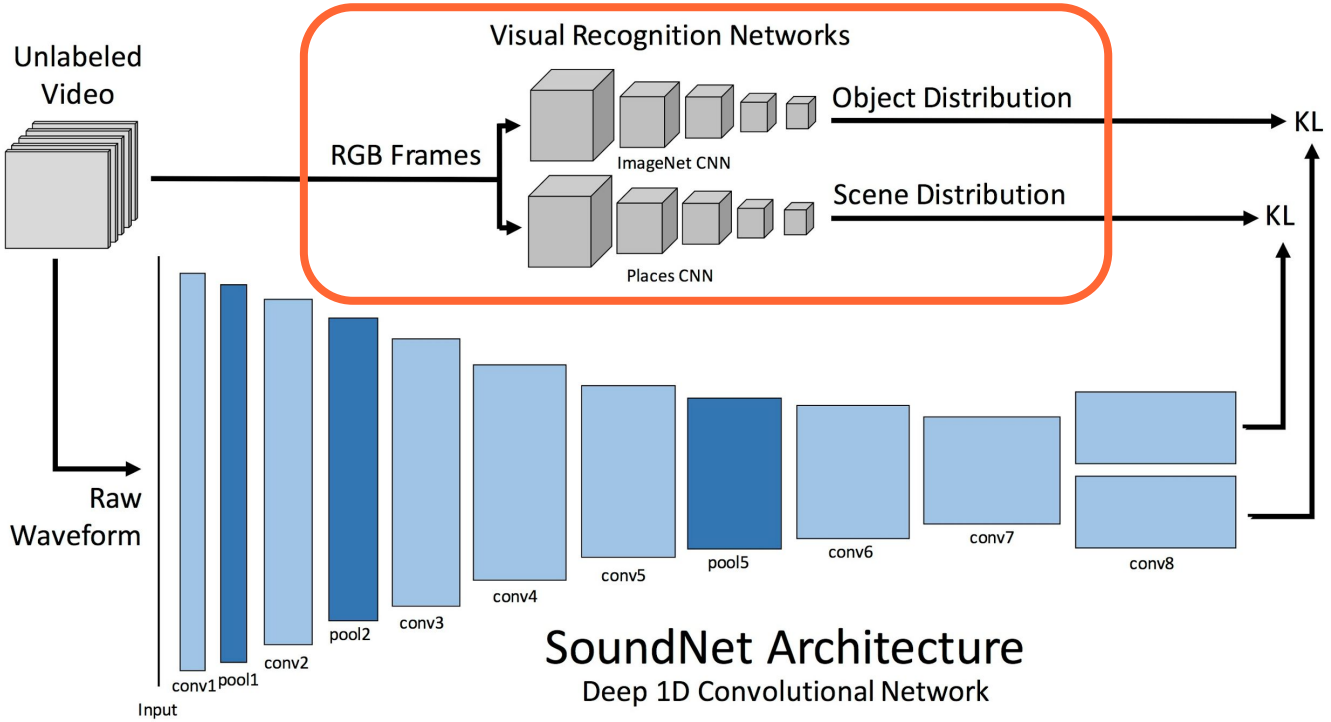
Pretrained visual ConvNets supervise the training of a model for sound representation



Aytar, Vondrick and Torralba. Soundnet: Learning sound representations from unlabeled video. NIPS 2016

Transfer Learning: SoundNet

Videos for training are unlabeled. Relies on Convnets trained on labeled images.



Aytar, Vondrick and Torralba. Soundnet: Learning sound representations from unlabeled video. NIPS 2016

Transfer Learning: SoundNet

prob. distribution of vision network

Kullback-Leibler Divergence: $D_{KL}(P||Q) = \sum_j P_j \log \frac{P_j}{Q_j}$

prob. distribution of sound network

Cross-Entropy: $$CE(P,Q) = -\sum_{j}^{N} P_j log(Q_j)$$

Aytar, Vondrick and Torralba. Soundnet: Learning sound representations from unlabeled video. NIPS 2016

Transfer Learning: SoundNet

Recognizing Objects and Scenes from Sound

Aytar, Vondrick and Torralba. Soundnet: Learning sound representations from unlabeled video. NIPS 2016

Transfer Learning: SoundNet

Recognizing Objects and Scenes from Sound

Aytar, Vondrick and Torralba. Soundnet: Learning sound representations from unlabeled video. NIPS 2016

Transfer Learning: SoundNet

Hidden layers of Soundnet are used to train a standard SVM classifier that outperforms state of the art.

Method	Accuracy
RG [29]	69%
LTT [21]	72%
RNH [30]	77%
Ensemble [34]	78%
SoundNet	**88%**

Table 3: **Acoustic Scene Classification on DCASE:** We evaluate classification accuracy on the DCASE dataset. By leveraging large amounts of unlabeled video, SoundNet generally outperforms hand-crafted features by 10%.

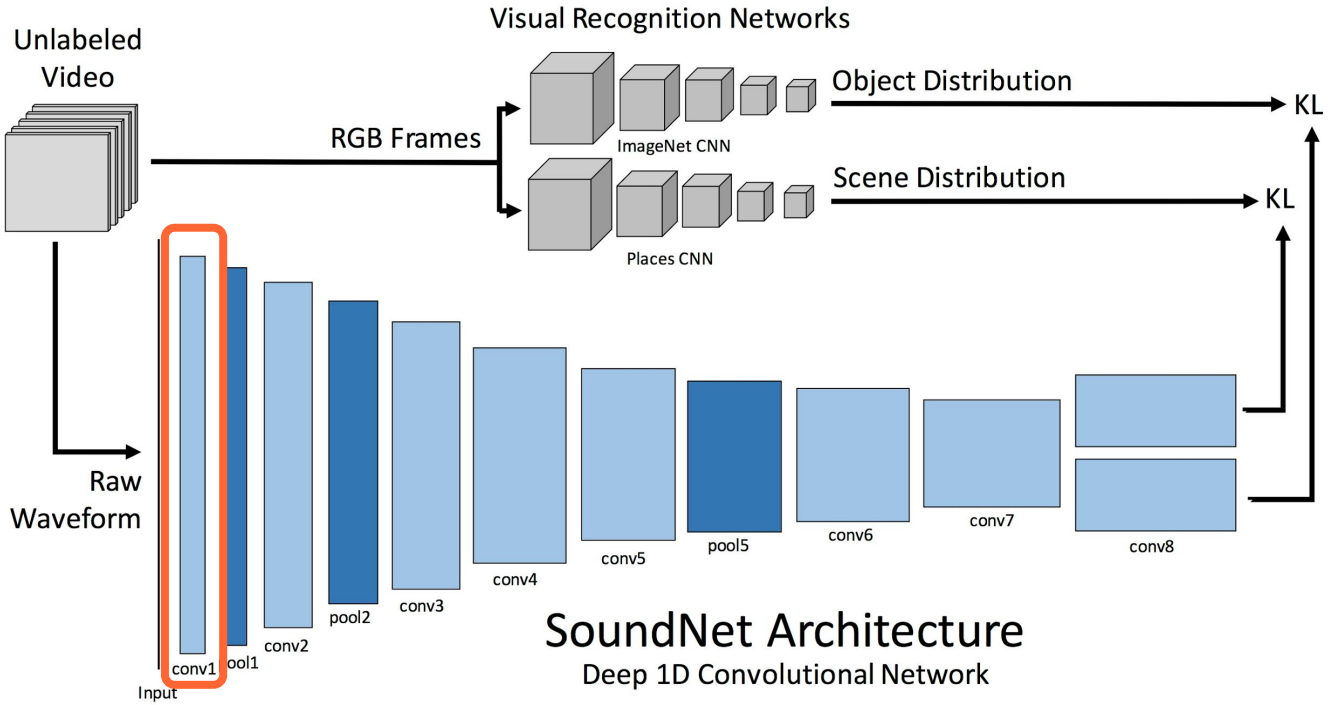
Method	Accuracy on ESC-50	Accuracy on ESC-10
SVM-MFCC [28]	39.6%	67.5%
Convolutional Autoencoder	39.9%	74.3%
Random Forest [28]	44.3%	72.7%
Piczak ConvNet [27]	64.5%	81.0%
SoundNet	**74.2%**	**92.2%**
Human Performance [28]	81.3%	95.7%

Table 4: **Acoustic Scene Classification on ESC-50 and ESC-10:** We evaluate classification accuracy on the ESC datasets. Results suggest that deep convolutional sound networks trained with visual supervision on unlabeled data outperforms baselines.

Aytar, Vondrick and Torralba. Soundnet: Learning sound representations from unlabeled video. NIPS 2016

Transfer Learning: SoundNet

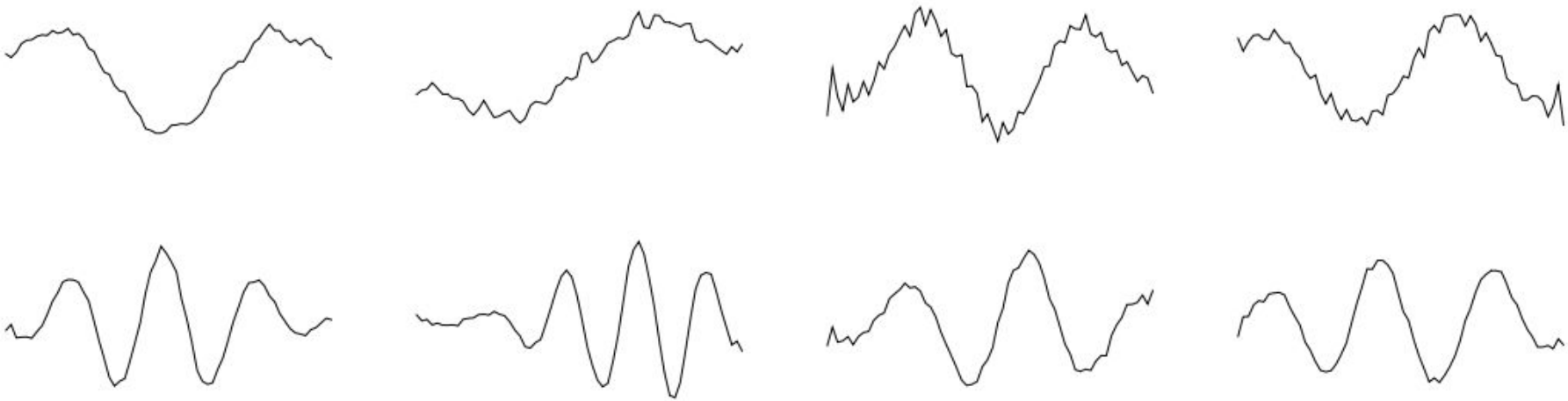
Visualization of the 1D filters over raw audio in conv1.



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016.

Transfer Learning: SoundNet

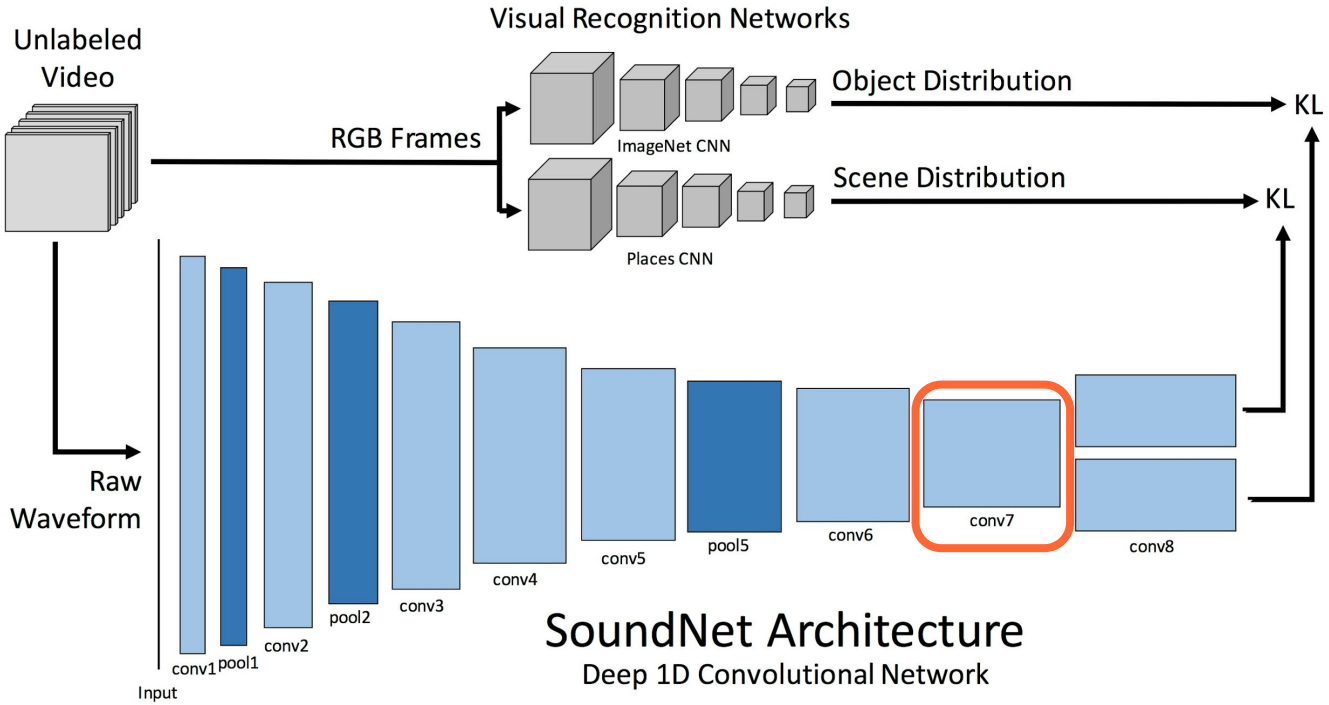
Visualization of the 1D filters over raw audio in conv1.



Aytar, Vondrick and Torralba. Soundnet: Learning sound representations from unlabeled video. NIPS 2016

Transfer Learning: SoundNet

Visualize samples that mostly activate a neuron in a late layer (conv7)



Aytar, Vondrick and Torralba. Soundnet: Learning sound representations from unlabeled video. NIPS 2016

Transfer Learning: SoundNet

Visualization of the video frames associated to the sounds that activate some of the last hidden units (conv7):

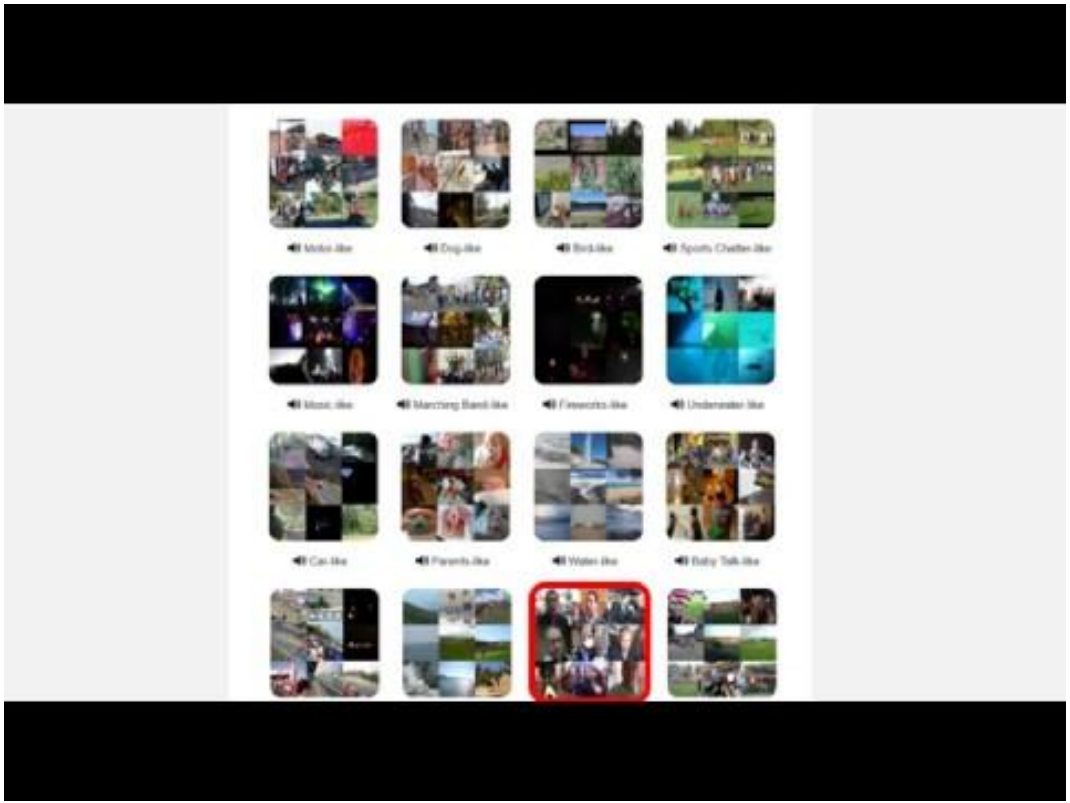
Baby Talk

Bubbles

Aytar, Vondrick and Torralba. Soundnet: Learning sound representations from unlabeled video. NIPS 2016

Transfer Learning: SoundNet

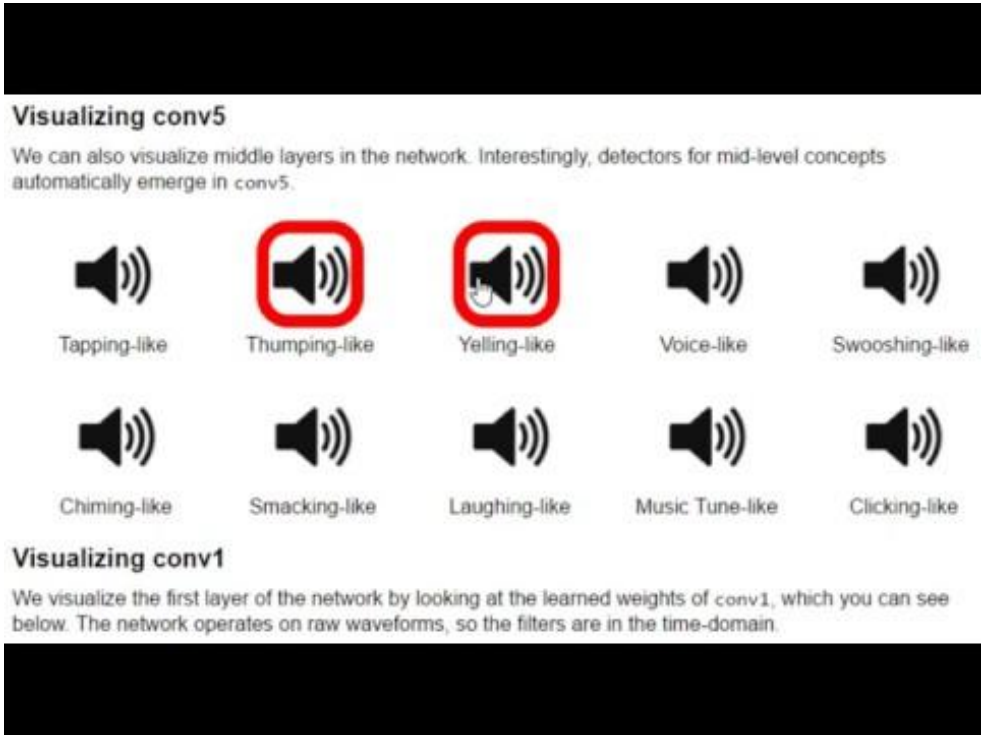
Hearing sounds that most activate a neuron in the sound network (conv7)



Aytar, Vondrick and Torralba. Soundnet: Learning sound representations from unlabeled video. NIPS 2016

Transfer Learning: SoundNet

Hearing sounds that most activate a neuron in the sound network (conv5)



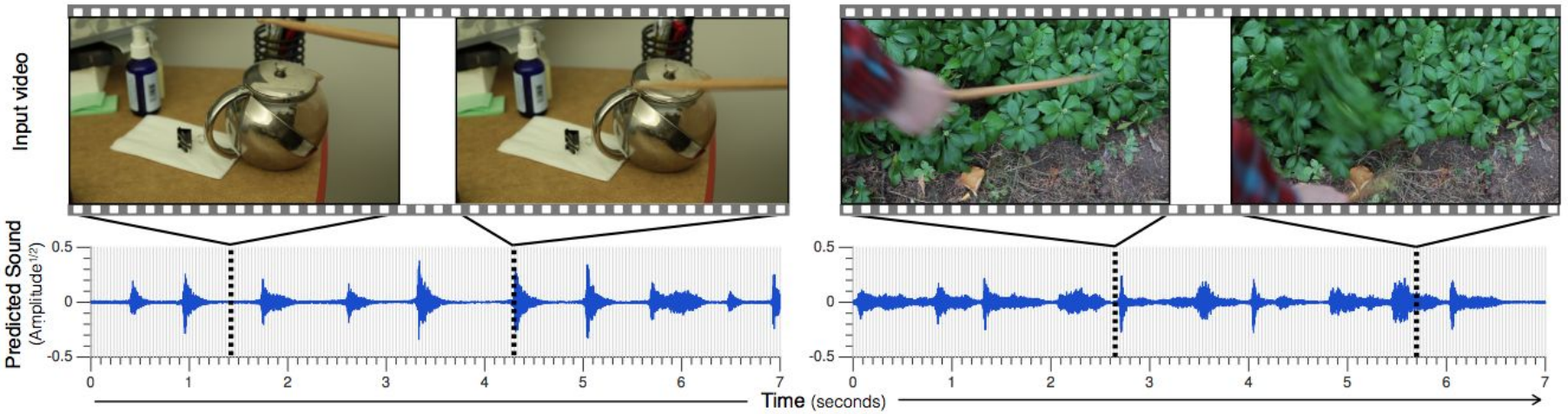
Aytar, Vondrick and Torralba. Soundnet: Learning sound representations from unlabeled video. NIPS 2016

Audio & Vision

- Transfer Learning
- **Sonorization**
- Speech generation
- Lip Reading

Sonorization

Learn synthesized sounds from videos of people hitting objects with a drumstick.

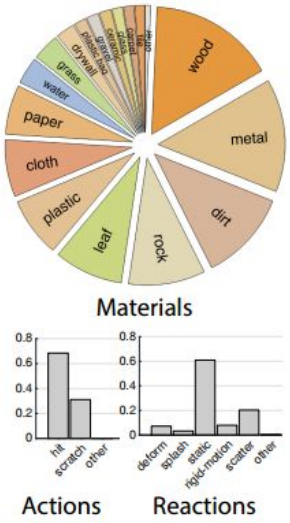


Are features learned when training for sound prediction useful to identify the object's material?

Owens et al. Visually indicated sounds. CVPR 2016

Sonorization

The Greatest Hits Dataset



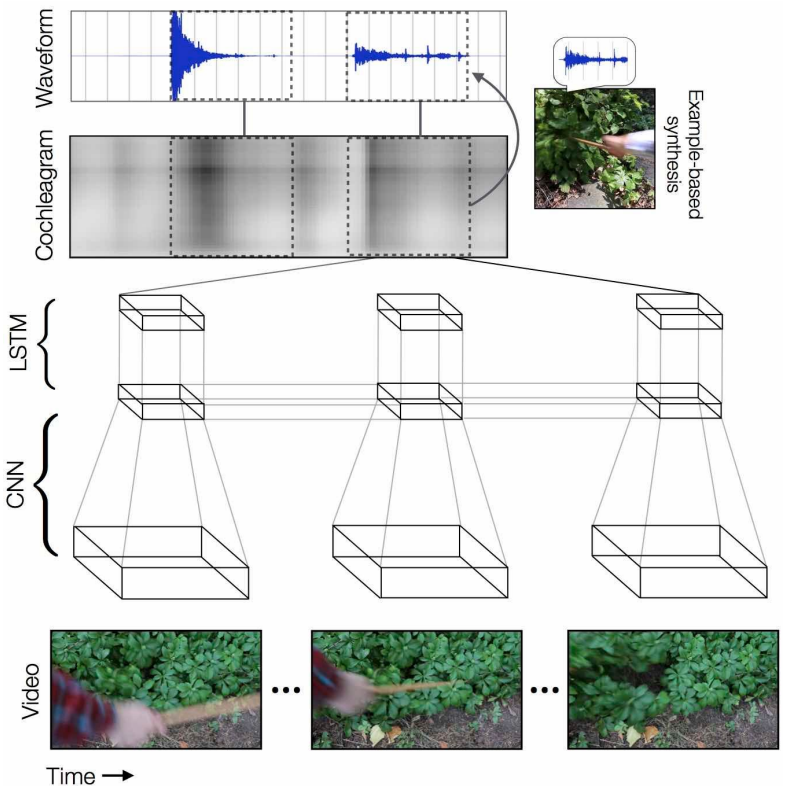
Owens et al. Visually indicated sounds. CVPR 2016

Sonorization

The Greatest Hits Dataset

Owens et al. Visually indicated sounds. CVPR 2016

Sonorization



- Map video sequences to sequences of sound features

$$E(\{\vec{s}_t\}) = \sum_{t=1}^{T} \rho(\|\vec{s}_t - \tilde{\vec{s}}_t\|_2),$$

- Sound features in the output are converted into waveform by example-based synthesis (i.e. nearest neighbor retrieval).

Owens et al. Visually indicated sounds. CVPR 2016

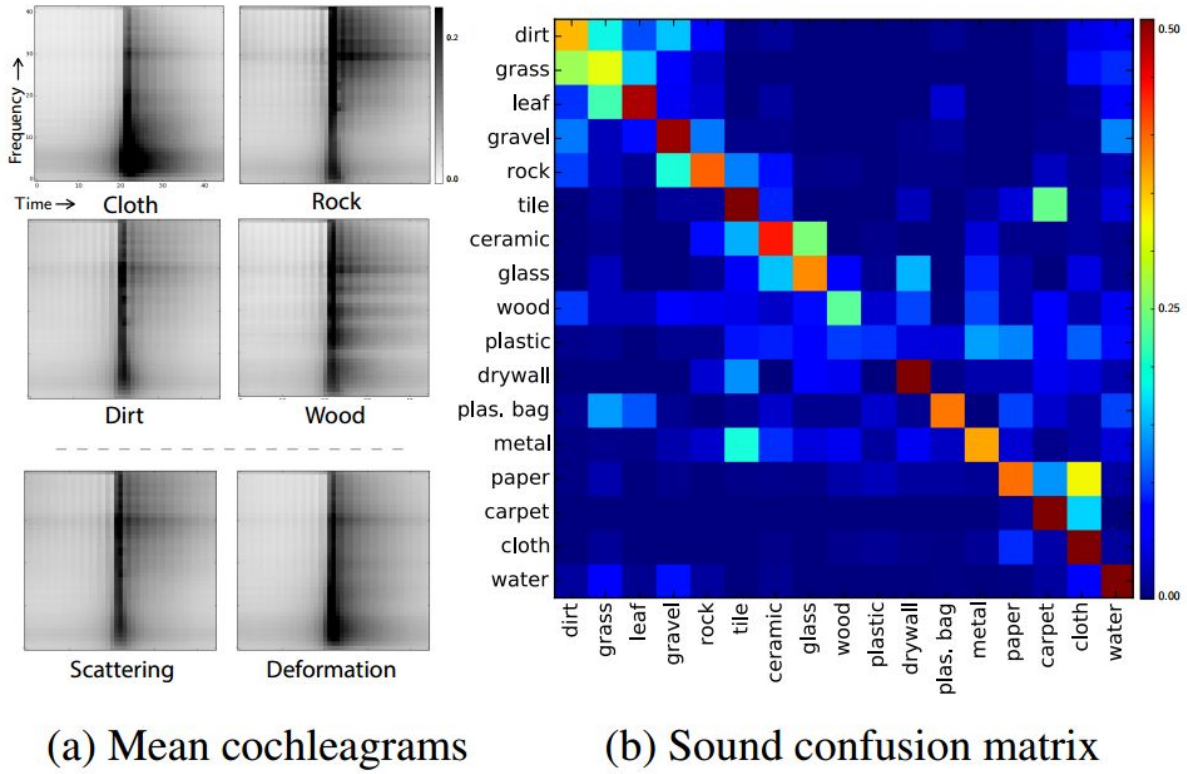
Sonorization

Owens et al. Visually indicated sounds. CVPR 2016

Sonorization

Owens et al. Visually indicated sounds. CVPR 2016

Sonorization



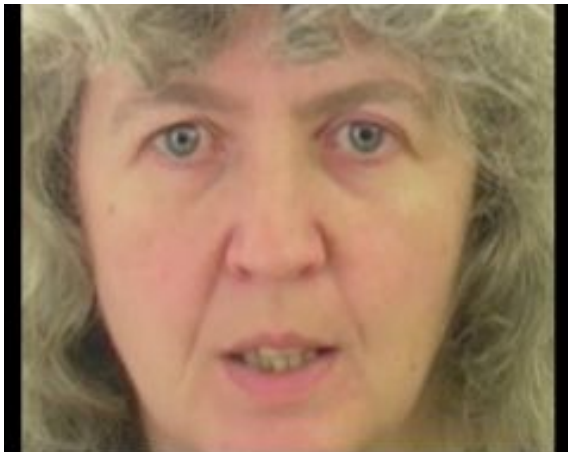
(a) Mean cochleagrams (b) Sound confusion matrix

Owens et al. Visually indicated sounds. CVPR 2016

Audio & Vision

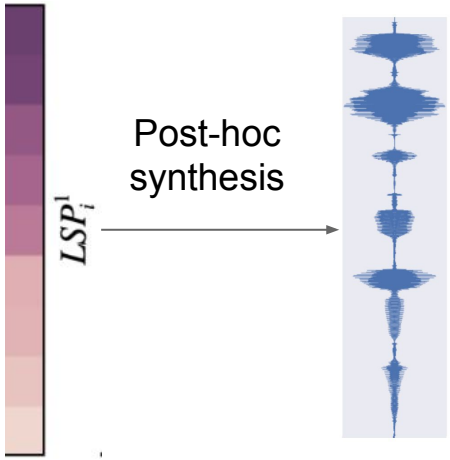
- Transfer Learning
- Sonorization
- **Speech generation**
- Lip Reading

Speech Generation from Video



Frame from a silent video

CNN (VGG)

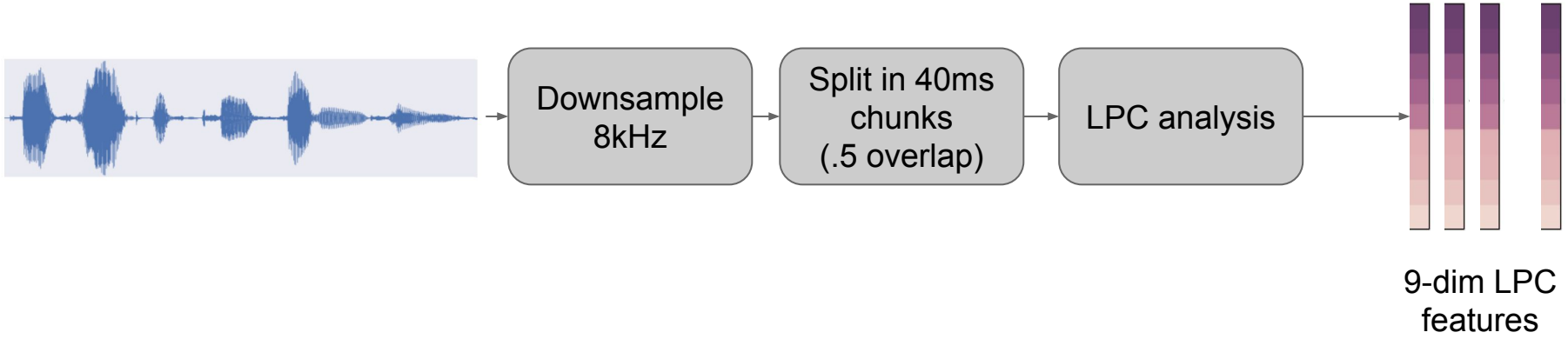


Audio feature

Ephrat et al. Vid2speech: Speech Reconstruction from Silent Video. *ICASSP 2017*

Speech Generation from Video

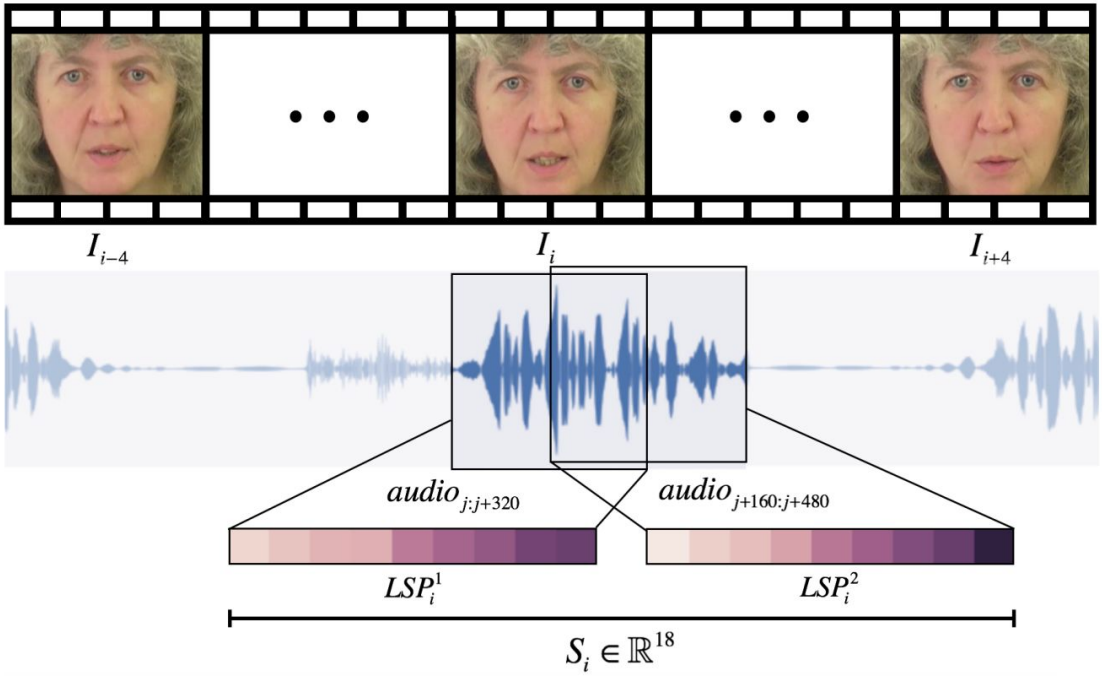
Waveform processing



Ephrat et al. Vid2speech: Speech Reconstruction from Silent Video. *ICASSP 2017*

Speech Generation from Video

Waveform processing (2 LSP features/ frame)



Ephrat et al. Vid2speech: Speech Reconstruction from Silent Video. *ICASSP 2017*

Speech Generation from Video

GRID Dataset

Command	Color	Preposition	Letter	Digit	Adverb
bin	blue	at	A-Z	0-9	again
lay	green	by	minus W		now
place	red	in			please
set	white	with			soon

Cooke et al. An audio-visual corpus for speech perception and automatic speech recognition. Journal of the Acoustical Society of America, 2006

Speech Generation from Video

Ephrat et al. Vid2speech: Speech Reconstruction from Silent Video. *ICASSP 2017*

Audio & Vision

- Transfer Learning
- Sonorization
- Speech generation
- **Lip Reading**

Lip Reading: LipNet

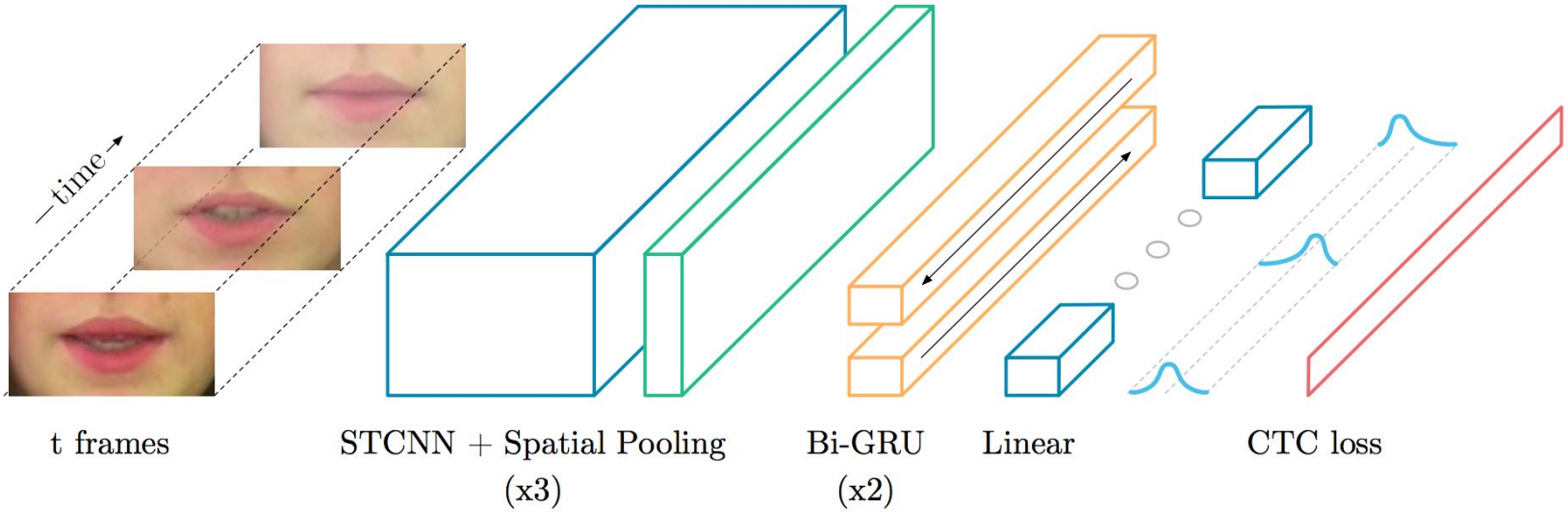


Figure 1: LipNet architecture. A sequence of T frames is used as input, and is processed by 3 layers of STCNN, each followed by a spatial max-pooling layer. The features extracted are processed by 2 Bi-GRUs; each time-step of the GRU output is processed by a linear layer and a softmax. This end-to-end model is trained with CTC.

Assael et al. LipNet: Sentence-level Lipreading. *arXiv Nov 2016*

Lip Reading: LipNet

Input (frames) and output (sentence) sequences are not aligned

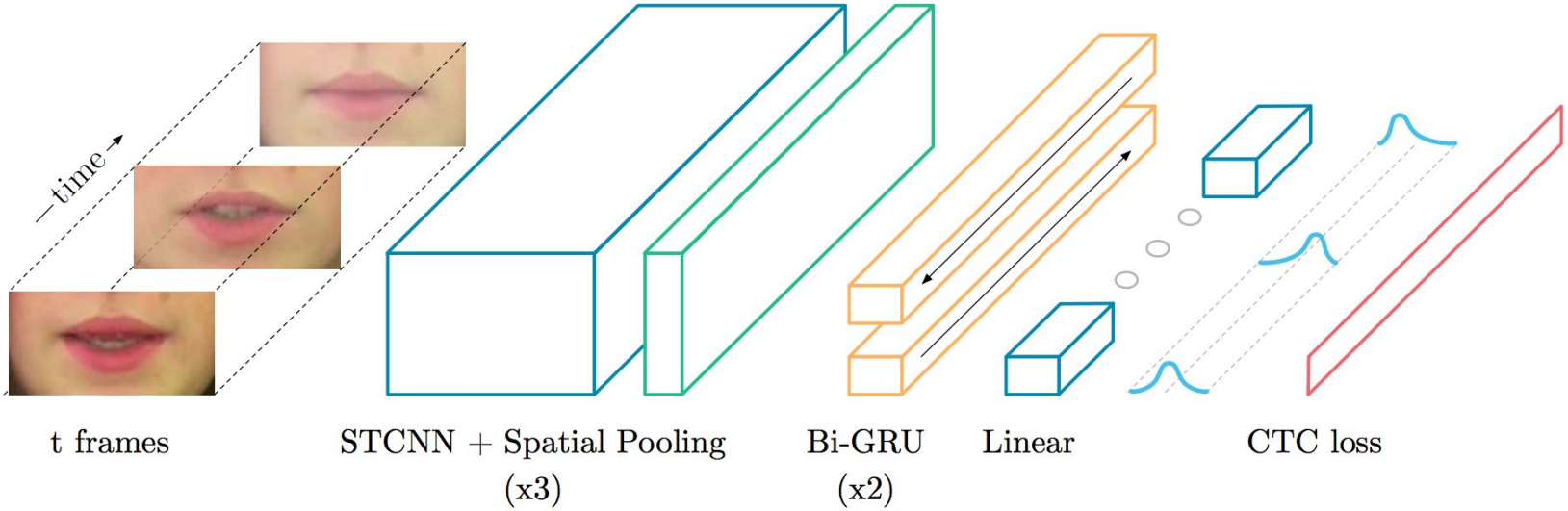


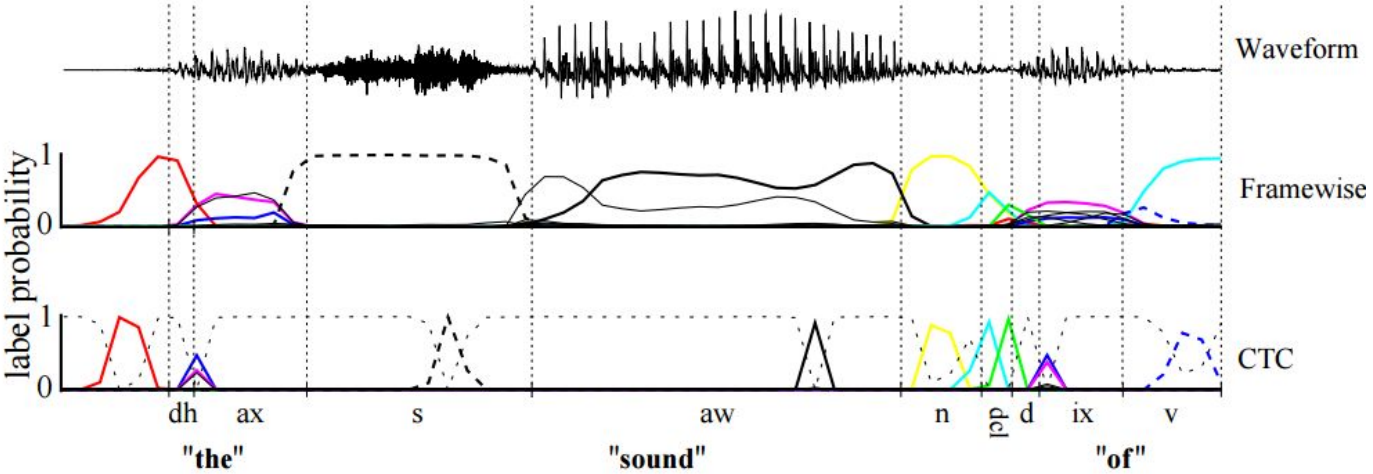
Figure 1: LipNet architecture. A sequence of T frames is used as input, and is processed by 3 layers of STCNN, each followed by a spatial max-pooling layer. The features extracted are processed by 2 Bi-GRUs; each time-step of the GRU output is processed by a linear layer and a softmax. This end-to-end model is trained with CTC.

Assael et al. LipNet: Sentence-level Lipreading. *arXiv Nov 2016*

Lip Reading: LipNet

CTC Loss: Connectionist temporal classification

- Avoiding the need for alignment between input and output sequence by predicting an additional “_” blank word
- Before computing the loss, repeated words and blank tokens are removed
- “a _ a b _ ” == “_ a a _ _ b b” == “a a b”



Graves et al. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. ICML 2006

Lip Reading: LipNet

Assael et al. [LipNet: Sentence-level Lipreading.](#) *arXiv Nov 2016*

Lip Reading: Watch, Listen, Attend & Spell

Figure 3. **Top:** Original still images from the BBC lip reading dataset – News, Question Time, Breakfast, Newsnight (from left to right). **Bottom:** The mouth motions for 'afternoon' from two different speakers. The network sees the areas inside the red squares.

Set	Dates	# Utter.	Vocab
Train	01/2010 - 12/2015	101,195	16,501
Val	01/2016 - 02/2016	5,138	4,572
Test	03/2016 - 09/2016	11,783	6,882
All		118,116	17,428

Chung et al. Lip reading sentences in the wild. *arXiv Nov 2016*

Lip Reading: Watch, Listen, Attend & Spell

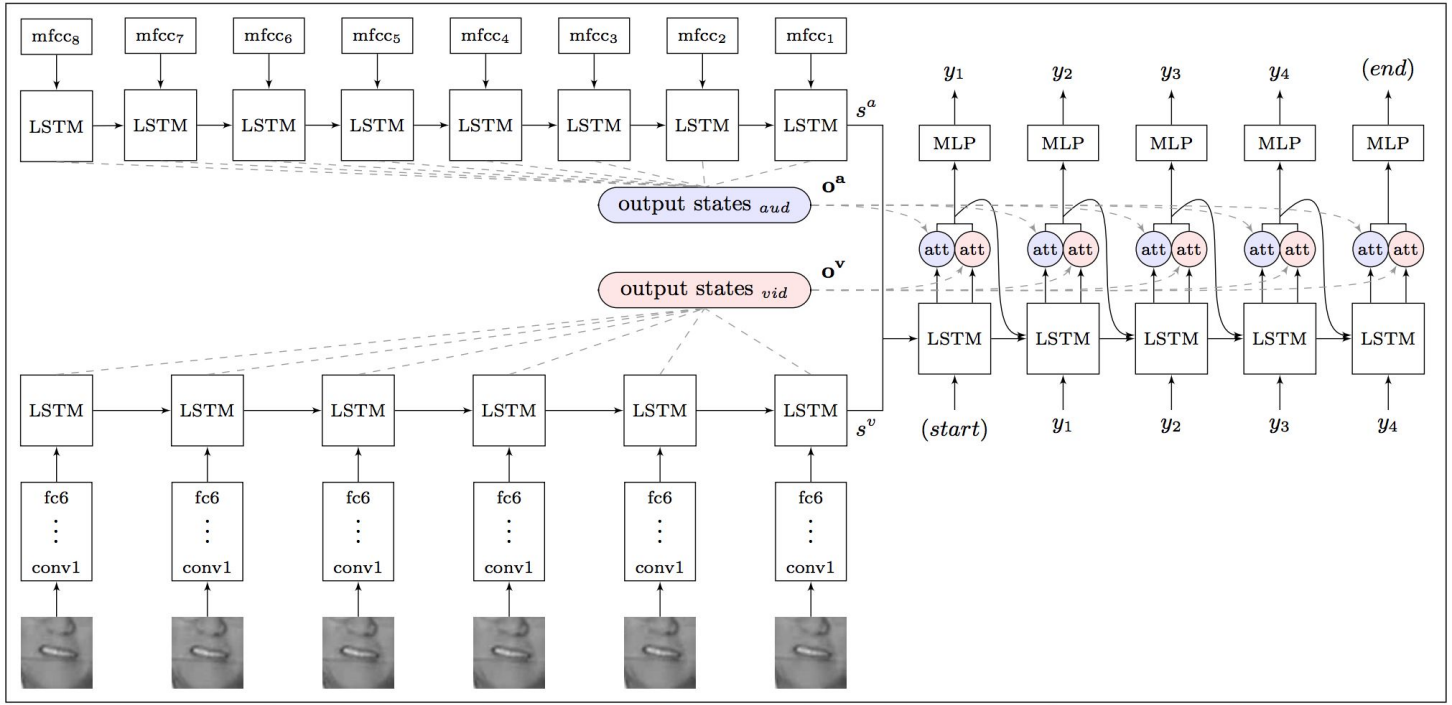


Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character y_i, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

Chung et al. Lip reading sentences in the wild. *arXiv Nov 2016*

Lip Reading: Watch, Listen, Attend & Spell

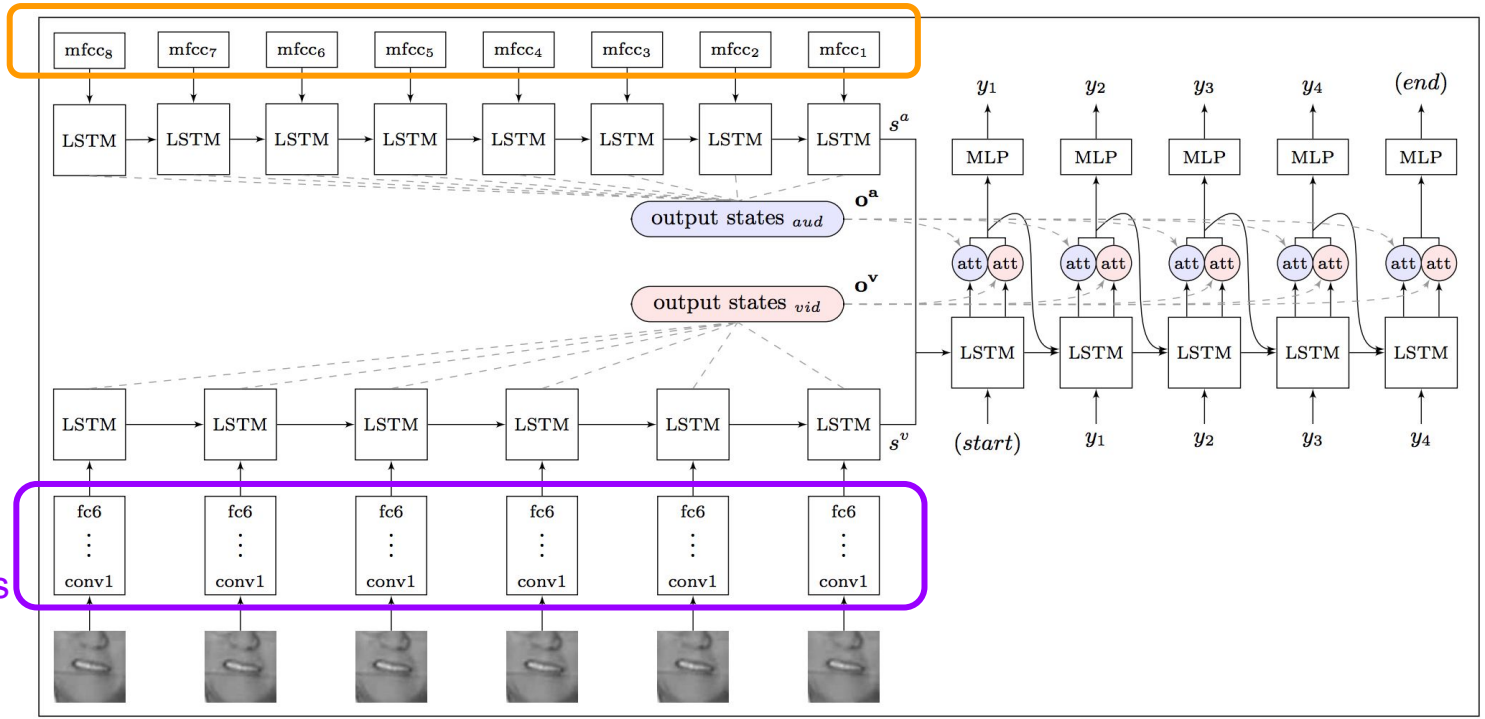


Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character y_i, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

Chung et al. Lip reading sentences in the wild. *arXiv Nov 2016*

Lip Reading: Watch, Listen, Attend & Spell

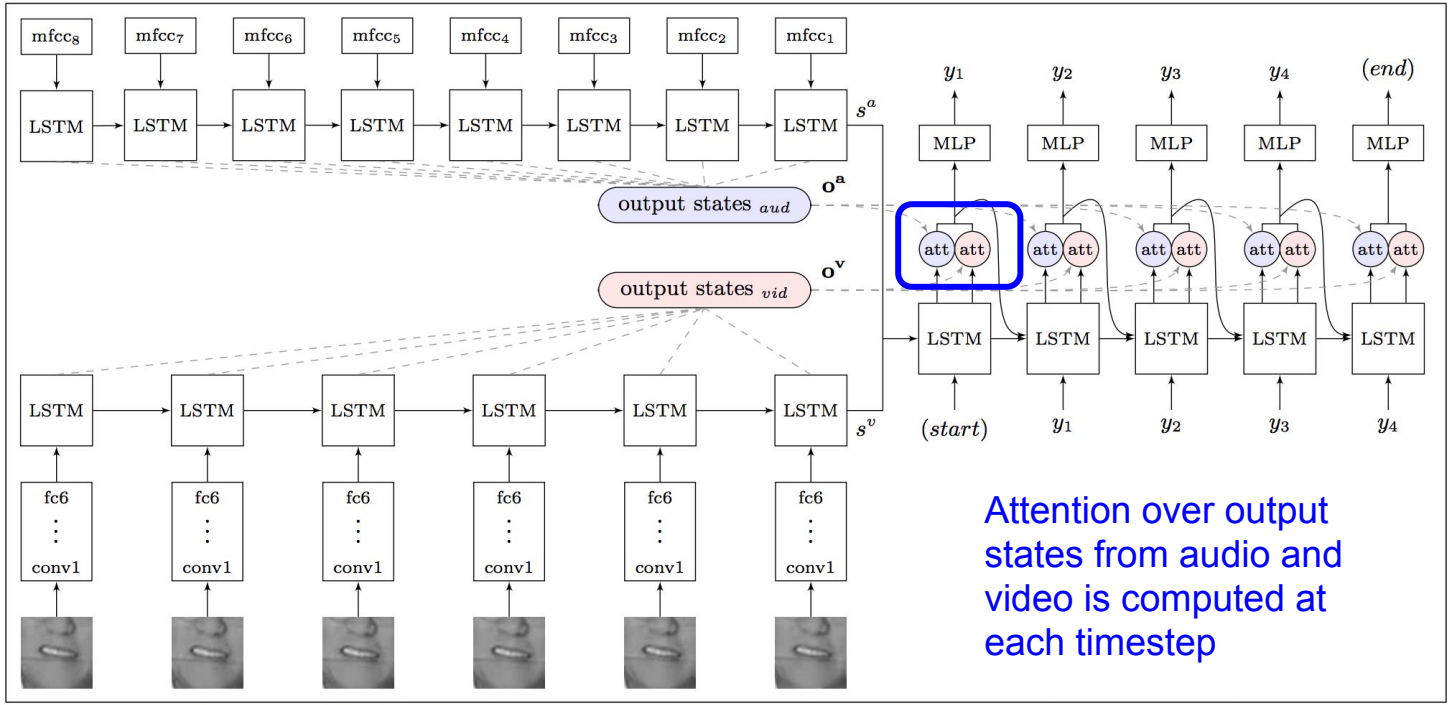


Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character y_i, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

Chung et al. Lip reading sentences in the wild. *arXiv Nov 2016*

Lip Reading: Watch, Listen, Attend & Spell

Method	SNR	CER	WER	BLEU†
Lips only				
Professional‡	-	58.7%	73.8%	23.8
WAS	-	59.9%	76.5%	35.6
WAS+CL	-	47.1%	61.1%	46.9
WAS+CL+SS	-	42.4%	58.1%	50.0
WAS+CL+SS+BS	-	39.5%	50.2%	54.9
Audio only				
Google Speech API	clean	17.6%	22.6%	78.4
Kaldi SGMM+MMI⋆	clean	9.7%	16.8%	83.6
LAS+CL+SS+BS	clean	10.4%	17.7%	84.0
LAS+CL+SS+BS	10dB	26.2%	37.6%	66.4
LAS+CL+SS+BS	0dB	50.3%	62.9%	44.6
Audio and lips				
WLAS+CL+SS+BS	clean	7.9%	13.9%	87.4
WLAS+CL+SS+BS	10dB	17.6%	27.6%	75.3
WLAS+CL+SS+BS	0dB	29.8%	42.0%	63.1

Chung et al. Lip reading sentences in the wild. *arXiv Nov 2016*

Lip Reading: Watch, Listen, Attend & Spell

LipNet

Methods	LRW [9]	GRID [11]
Lan *et al.* [23]	-	35.0%
Wand *et al.* [39]	-	20.4%
Assael *et al.* [2]	-	4.8%
Chung and Zisserman [9]	38.9%	-
WAS (ours)	**23.8%**	**3.0%**

Table 8. **Word error rates** on external lip reading datasets.

Chung et al. Lip reading sentences in the wild. *arXiv Nov 2016*

Lip Reading: Watch, Listen, Attend & Spell

Chung et al. Lip reading sentences in the wild. *arXiv Nov 2016*

Summary

- Transfer learning
- Sonorization
- Speech generation
- Lip Reading

Questions?

Multimodal Learning

	Feature Learning	Supervised Training	Testing
Classic Deep Learning	Audio	Audio	Audio
	Video	Video	Video
Multimodal Fusion	A + V	A + V	A + V
Cross Modality Learning	A + V	Video	Video
	A + V	Audio	Audio
Shared Representation Learning	A + V	Audio	Video
	A + V	Video	Audio

Ngiam et al. Multimodal Deep Learning. ICML 2011

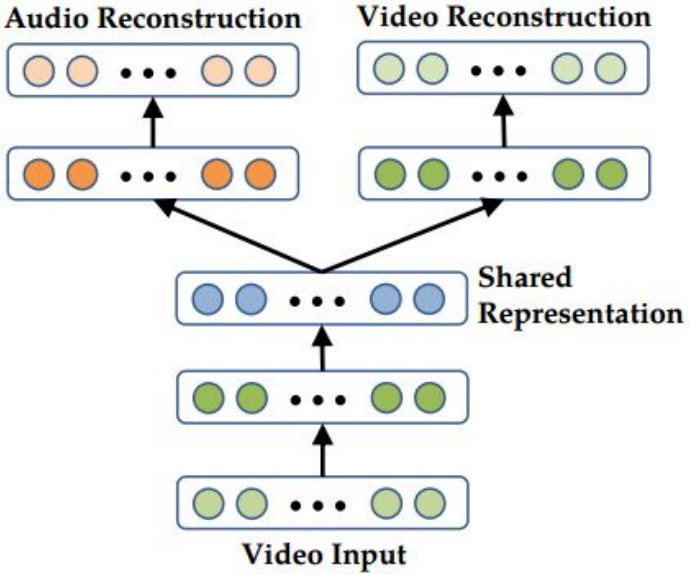
Multimodal Learning

	Feature Learning	Supervised Training	Testing
Classic Deep Learning	Audio	Audio	Audio
	Video	Video	Video
Multimodal Fusion	A + V	A + V	A + V
Cross Modality Learning	A + V	Video	Video
	A + V	Audio	Audio
Shared Representation Learning	A + V	Audio	Video
	A + V	Video	Audio

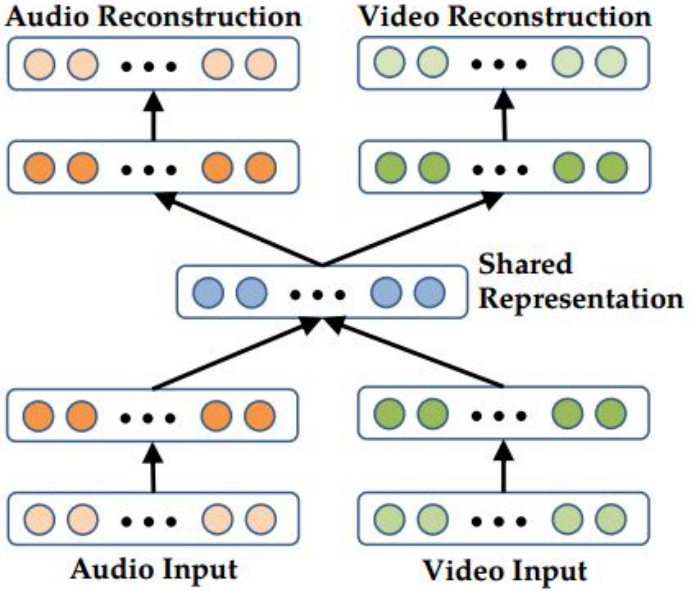
Ngiam et al. Multimodal Deep Learning. ICML 2011

Cross-Modality

Denoising autoencoder required to reconstruct both modalities given only one or both in the input



(a) Video-Only Deep Autoencoder



(b) Bimodal Deep Autoencoder

Ngiam et al. Multimodal Deep Learning. ICML 2011

Cross-Modality

Video classification: Add Linear SVM on top of learned features

- Combination of audio&video as inputs gives relative improvement (on AVLetters)
- Using video as input to predict both audio& video gives SotA results in both

Feature Representation	Accuracy
Baseline Preprocessed Video	46.2%
RBM Video (Figure 2b)	54.2%±3.3%
Video-Only Deep Autoencoder (Figure 3a)	**64.4%**±2.4%
Bimodal Deep Autoencoder (Figure 3b)	59.2%
Multiscale Spatial Analysis (Matthews et al., 2002)	44.6%
Local Binary Pattern (Zhao & Barnard, 2009)	58.85%

(a) AVLetters

Feature Representation	Accuracy
Baseline Preprocessed Video	58.5%
RBM Video (Figure 2b)	65.4%±0.6%
Video-Only Deep Autoencoder (Figure 3a)	**68.7%**±1.8%
Bimodal Deep Autoencoder (Figure 3b)	66.7%
Discrete Cosine Transform (Gurban & Thiran, 2009)	64% †§
Active Appearence Model (Papandreou et al., 2007)	75.7% †
Active Appearence Model (Pitsikalis et al., 2006)	68.7% †
Fused Holistic+Patch (Lucey & Sridharan, 2006)	77.08% †
Visemic AAM (Papandreou et al., 2009)	83% †§

(b) CUAVE Video

Ngiam et al. Multimodal Deep Learning. ICML 2011

Cross-Modality

Audio classification: Add Linear SVM on top of learned features

- Performance of cross-modal features below SotA
- Still, remarkable performance when using video as input to classify audio

Feature Representation	Accuracy (Clean Audio)	Accuracy (Noisy Audio)
(a) Audio RBM (Figure 2a)	**95.8%**	75.8% ± 2.0%
(b) Video-only Deep Autoencoder (Figure 3a)	68.7%	68.7%
(c) Bimodal Deep Autoencoder (Figure 3b)	90.0%	77.3% ± 1.4%
(d) **Bimodal + Audio RBM**	94.4%	**82.2%** ± 1.2%
(e) Video-only Deep AE + Audio-RBM	87.0%	76.6% ± 0.8%

Ngiam et al. Multimodal Deep Learning. ICML 2011

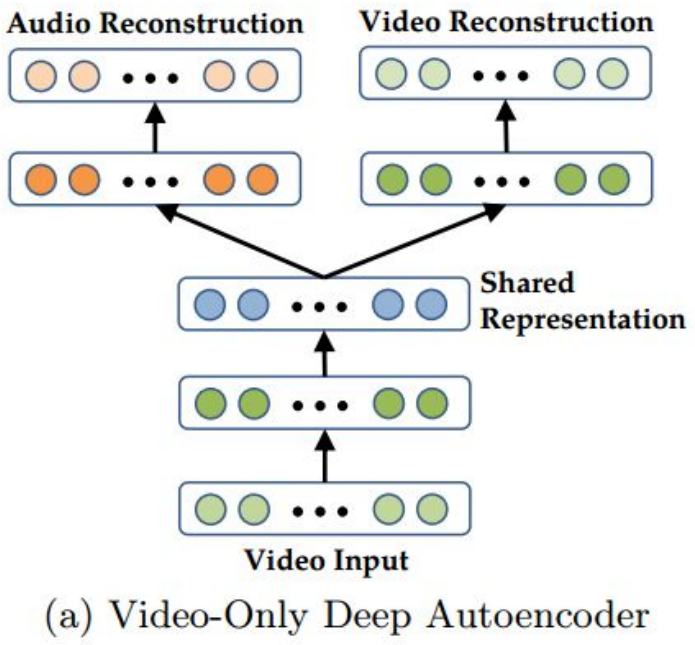
Multimodal Learning

	Feature Learning	Supervised Training	Testing
Classic Deep Learning	Audio	Audio	Audio
	Video	Video	Video
Multimodal Fusion	A + V	A + V	A + V
Cross Modality Learning	A + V	Video	Video
	A + V	Audio	Audio
Shared Representation Learning	A + V	Audio	Video
	A + V	Video	Audio

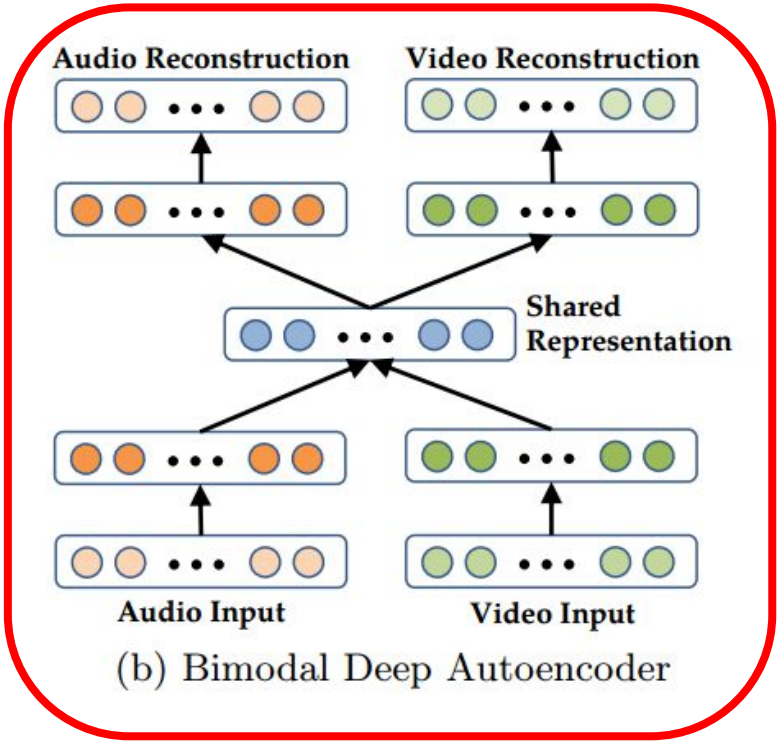
Ngiam et al. Multimodal Deep Learning. ICML 2011

Shared Learning

First, learn a shared representation using the two modalities as input



(a) Video-Only Deep Autoencoder

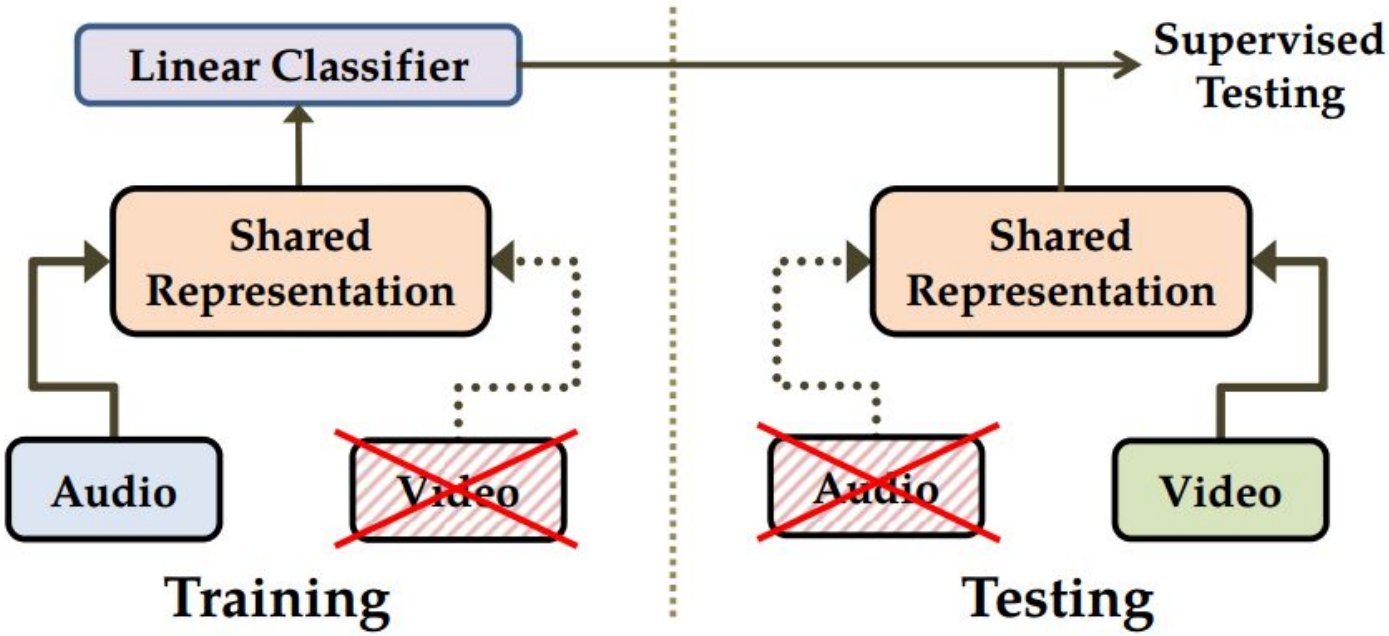


(b) Bimodal Deep Autoencoder

Ngiam et al. Multimodal Deep Learning. ICML 2011

Shared Learning

Then, train linear classifiers on top of shared representations (previously trained on audio/video data) using only one of the two modalities as input. Testing is performed with the other modality.



Ngiam et al. Multimodal Deep Learning. ICML 2011

Shared Learning

Train Linear classifiers on top of shared representations (previously trained on audio/video data) using only one of the two modalities as input. Testing is performed with the other modality.

Table 4: Shared representation learning on CUAVE. Chance performance is at 10%.

Train/Test	Method	Accuracy
Audio/Video	Raw-CCA	41.9%
	RBM-CCA Features	**57.3%**
	Bimodal Deep AE	30.7%
Video/Audio	Raw-CCA	42.9%
	RBM-CCA Features	**91.7%**
	Bimodal Deep AE	24.3%

Ngiam et al. Multimodal Deep Learning. ICML 2011