Modelling Acoustic-Feature Dependencies with Artificial Neural Networks: Trajectory-RNADE

Benigno Uria, Iain Murray,
Steve Renals, Cassia Valentini

University of Edinburgh

Motivation


Text
Front-end
Phonetic labels
Back-end
Acoustics
Vocoder
Waveform

Motivation

Usual parametric back-ends:

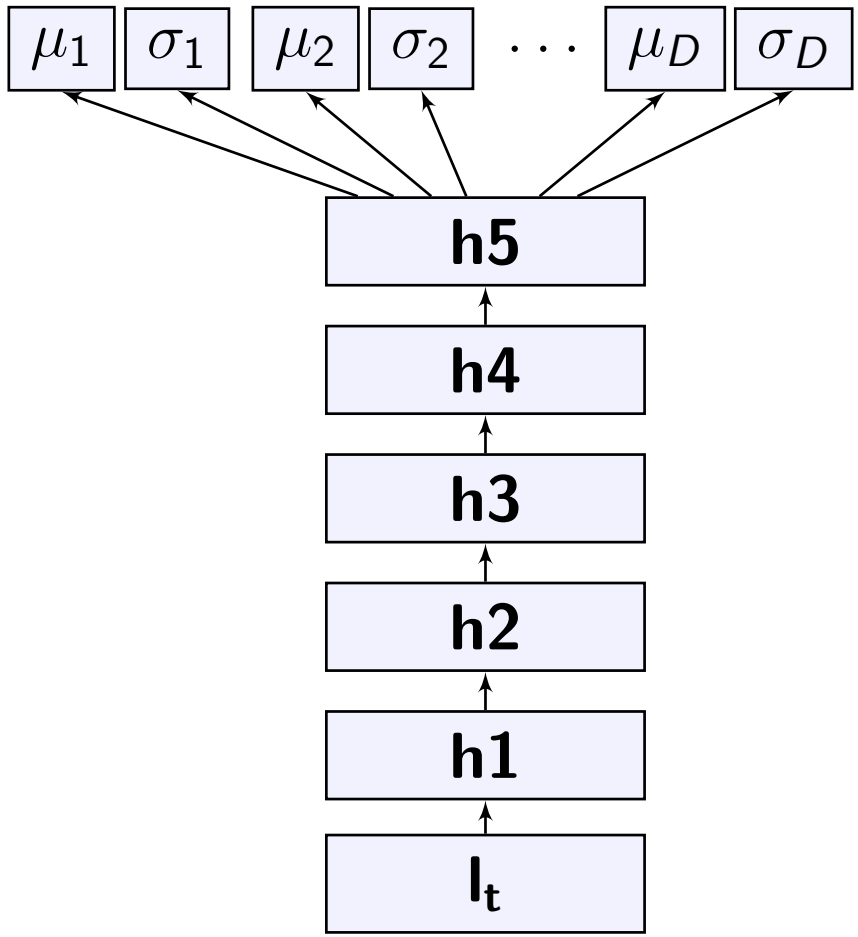
- Tied Gaussians or MoG (HTK)
- Regression ANN (fixed variance)
- Mixture Density Networks

Motivation

Usual parametric back-ends:

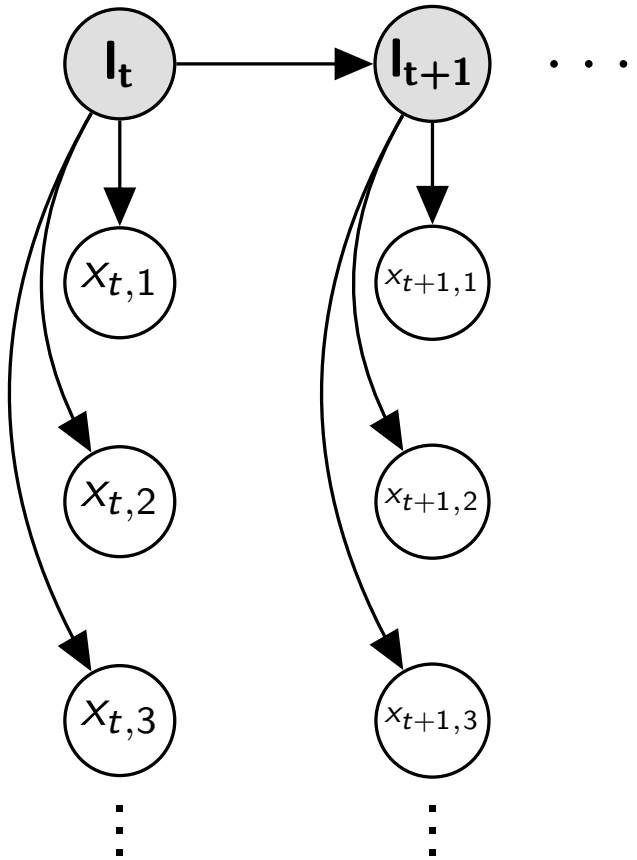
- Tied Gaussians or MoG (HTK)
- Regression ANN (fixed variance)
- **Mixture Density Networks**

Mixture density networks



Graphical model

HMM with Mixture Density Network conditionals:



$$p(\mathbf{x_t} \mid \mathbf{l_t}) = p(x_{t,1} \mid \mathbf{l_t}) p(x_{t,2} \mid \mathbf{l_t}) p(x_{t,3} \mid \mathbf{l_t}) \ldots$$

Why can't we sample?

Samples sound noisy.

Conditional independences given phonetic labels:

Why can't we sample?

Samples sound noisy.

Conditional independences given phonetic labels:

- Across time
    - Trajectory HMM formalism (later)

Why can't we sample?

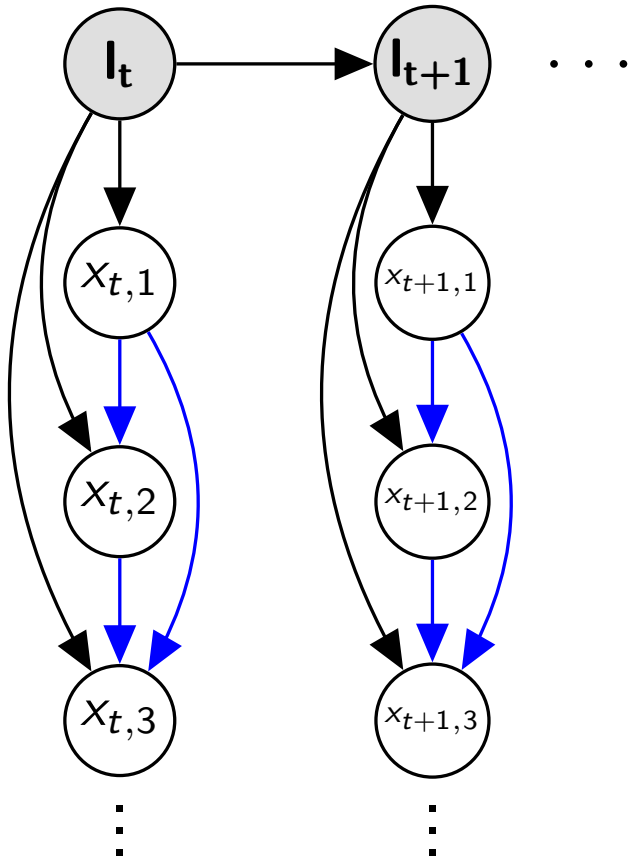
Samples sound noisy.

Conditional independences given phonetic labels:

- Across time
    - Trajectory HMM formalism (later)
- Across acoustic features
    - Henter *et al.*
    - Our focus

Acoustic feature dependencies

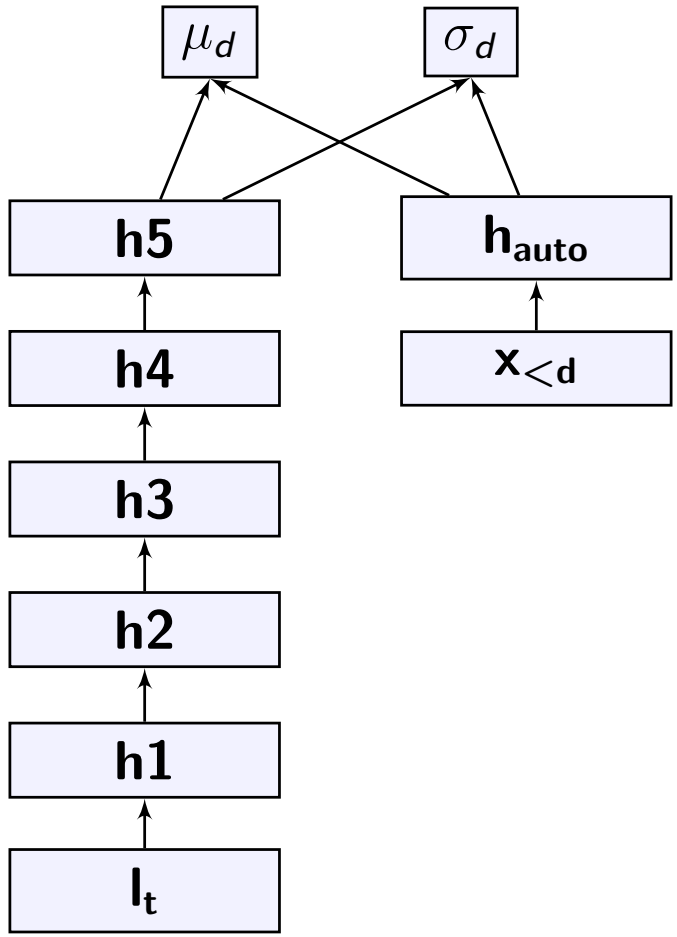
We propose an ANN that models:



$$p(\mathbf{x_t} \mid \mathbf{l_t}) = p(x_{t,1} \mid \mathbf{l_t}) p(x_{t,2} \mid x_{t,1}, \mathbf{l_t}) p(x_{t,3} \mid x_{t,1}, x_{t,2}, \mathbf{l_t}) \ldots$$

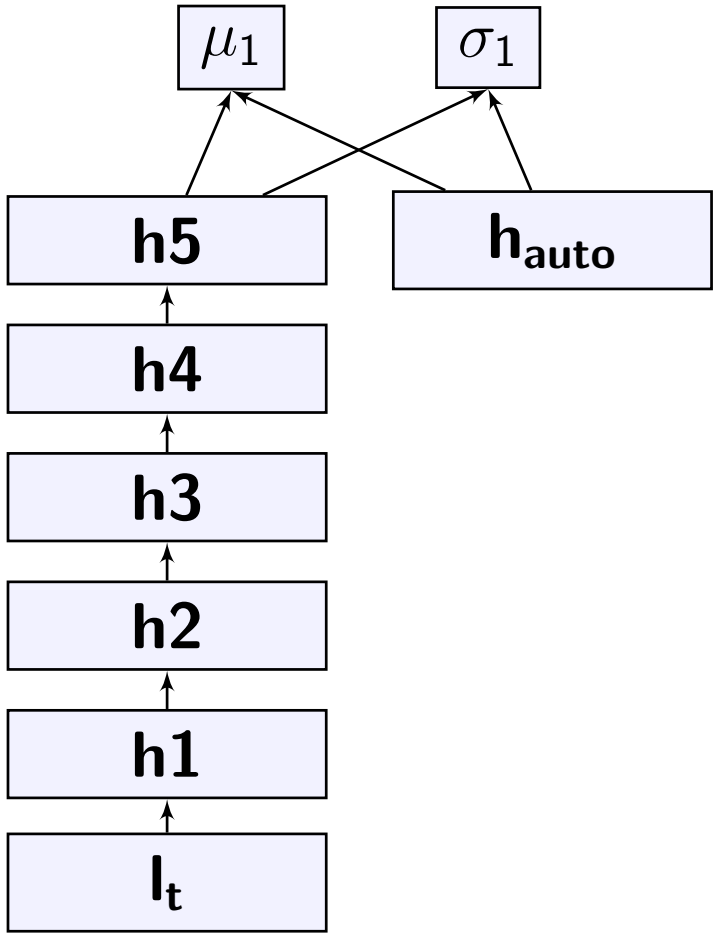
Conditional RNADE

RNADE = Real-valued Neural Autoregressive Density Estimator



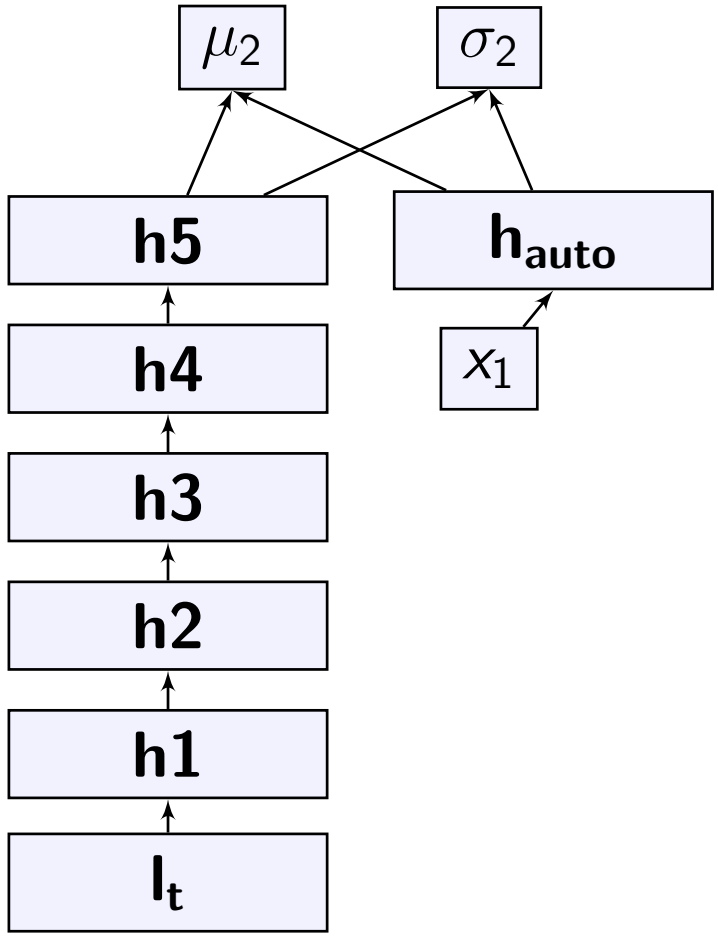
Conditional RNADE

How it works:



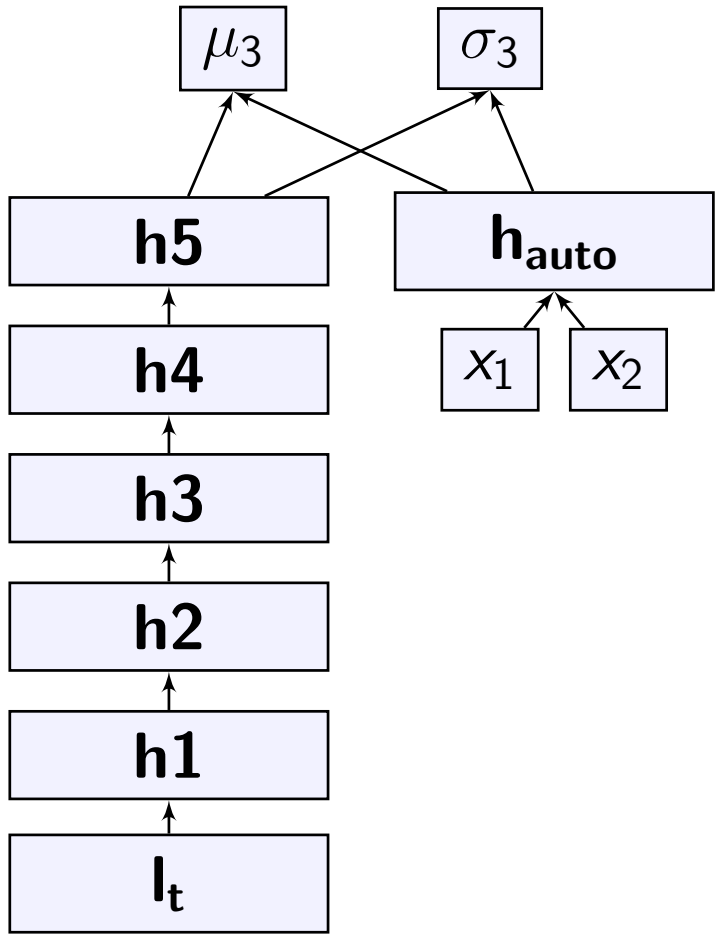
Conditional RNADE

How it works:



Conditional RNADE

How it works:



Conditional RNADE

- Can model non-linear dependencies
- Complexity $O(DH + H^2)$
- Tractable densities (unlike RBM, DBN, ...)
- Deep versions are slow ☹

Dependencies across time

Trajectory-HMM formalism

- Augment the vector with deltas (velocity and acceleration)
- Output a distribution for each frame
- Calculate distribution over constrained subset of trajectories

In RNADE:

- Output static and deltas simultaneously
- Predict the whole trajectory of a feature before the next feature

Experiments

- British male voice
- 2 hours of data
- STRAIGHT (v/uv, f_0, 60 melcep, 25 bap)

Subjective test results

- RNADE achieves much better likelihoods
- Forced preference test results:

MDN		RNADE	
Sample	Mean	Sample	Mean
19.4%	-	**80.6%**	-
-	33.6%	-	**66.4%**
-	-	16.0%	**84.0%**

- RNADE samples and means sound better
- Mean trajectories are still preferred

Future work

We still need better models

Trajectory-HMM formalism is limiting:

- Gaussian conditionals
- Wrong loss function
- Delay in generation (worse with NADE)

Alternatives:

- Fully autoregressive (time and features)
- Copula for dependencies across time

Questions

Papers and code available: `www.benignouria.com`