Acoustic Scene Classification from Few Examples

Ivan Bocharov, Tjalling Tjalkens and Bert de Vries

Eindhoven University of Technology, the Netherlands

Email: i.a.bocharov@tue.nl

EUSIPCO 2018, 05-09-2018

Use case/Problem statement

Images: wikipedia.org

Design constraints

- Ability to learn a new acoustic environment **from few** (preferably, a single) **examples**.

- An example is approximately **10-15 seconds long**.

- **Small computational footprint** is preferable.

Approach: probabilistic modeling

We use **probabilistic modeling approach**.

Model definition:

- Define a generative probabilistic model for acoustic scenes that contains classes c as latent variables:

$$p(x, z, c, \theta)$$

Training:

- Supervised training on a small in-situ recorded set of labeled waveforms

Classification:

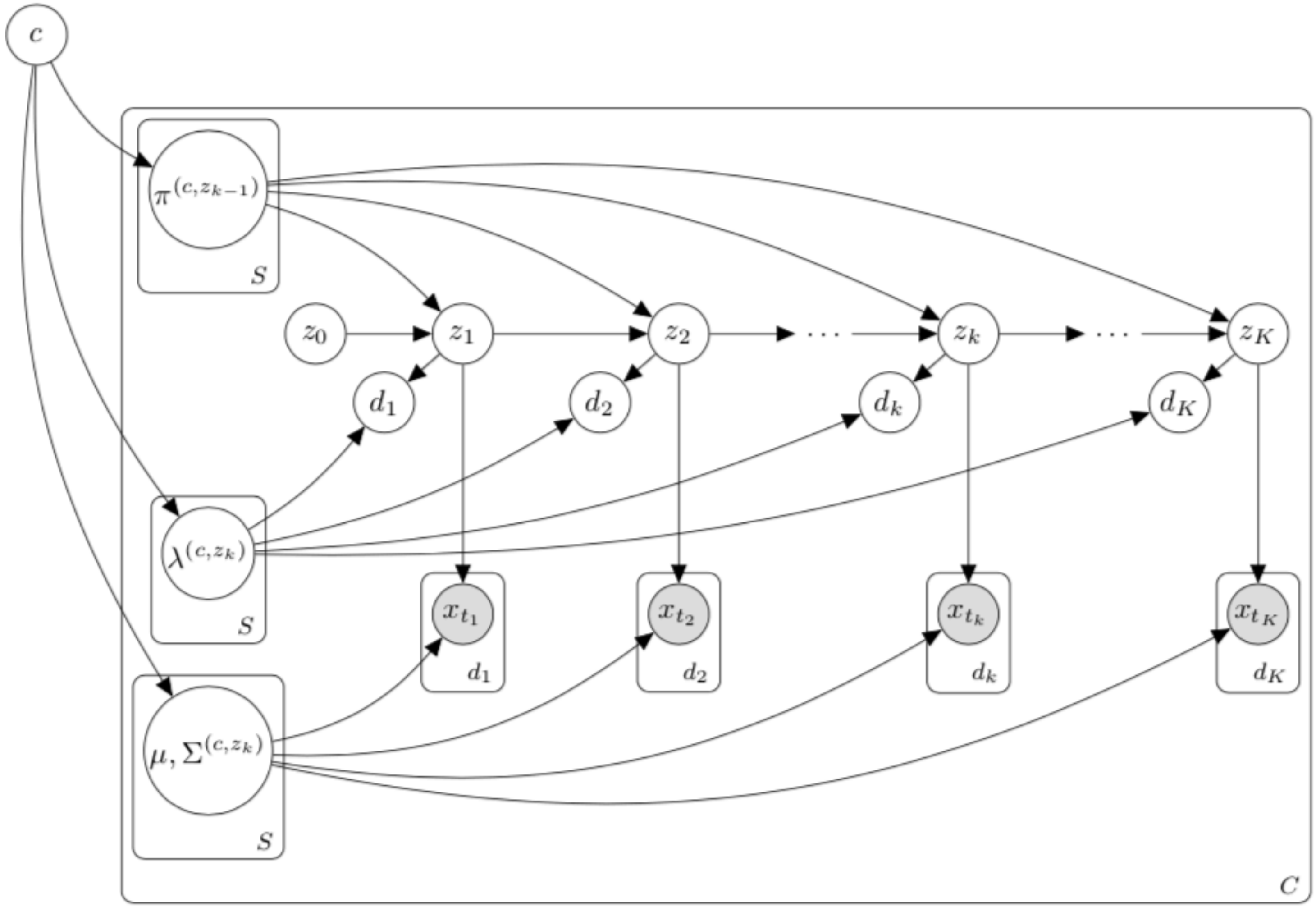
- Assign future streaming acoustic data to correct (or similar) classes

Approach: probabilistic modeling

Benefits:

- All tasks (learning, classification) can be formulated as inference tasks on generative model.
- Domain knowledge can be incorporated in a principled way via prior specification (parameters usually have explicit semantics).
- Structured approach; less data- and compute-greedy compared to deep learning alternatives.
- Steadily improving toolset (e.g. ForneyLab, Stan, Edward) allows for fast design iterations.

(Mixture of) Hidden Semi-Markov Model



(Mixture of) Hidden Semi-Markov Model

- Hierarchical organization:
  - Classes (~10 s)
  - States (~100 ms)
  - Observations (~1 ms)

- Explicit duration modeling allows to incorporate domain knowledge about state evolution.

- Small number of tunable parameters (fast inference).

Dataset

We use DCASE'2017 dataset for evaluation:

- 15 acoustic scenes
- ~65 min. of audio per class
- Wide variety of scenes (public transport, office, etc.)

Preprocessing:
extract 20 MFCC (40 ms window, 20 ms hop) + Δ + $\Delta\Delta$

Images: wikipedia.org





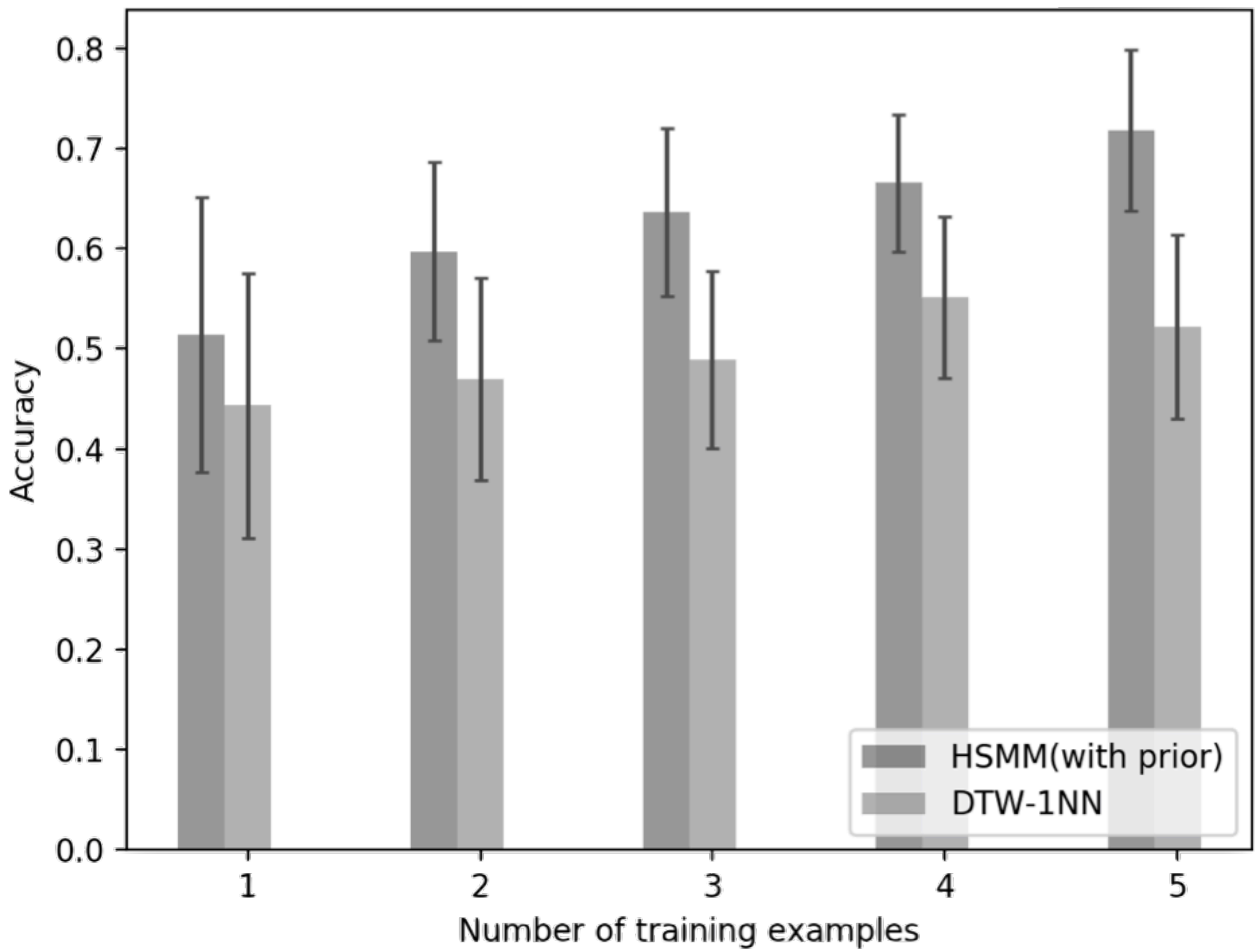

Evaluation protocol

To imitate our use case we use following protocol:

1. Randomly select 4 scenes from DCASE dataset.
2. Sample **M** training examples for each of scenes from development part of DCASE dataset.
3. Evaluation set is constructed by selecting all recordings for selected scenes from evaluation part of DCASE dataset.

We repeat this 20 times for each value of **M** in order to minimize the influence of random selection.

Results



Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1
2
3
4
5
Number of training examples
HSMM(with prior)
DTW-1NN

Summary and future work

- We present generative **probabilistic modeling approach** to **in-situ learning** of acoustic scene classifiers.
- Use case is **hearing aids personalization**, but applicable to other domains such as urban monitoring and elderly care.
- Specifically, we developed **HSMM-based acoustic scene classifier** that can be trained on very few (a single) recordings.
- **Priors** are not learned from the data and in principle **could (should) be learned** from the whole dataset in an offline unsupervised manner.
- Labels provided by users might not represent separate acoustic scenes.

Acknowledgments

We gratefully acknowledge Matthew Johnson and other developers of **pyhsmm** package:

https://github.com/mattjj/pyhsmm

Thank you