# K-shot Learning of Acoustic Context

**Ivan Bocharov***, **Bert de Vries***,† **and Tjalling Tjalkens***
*Eindhoven University of Technology, †GN Hearing BV, the Netherlands
{i.a.bocharov, bert.de.vries, t.j.tjalkens}@tue.nl

## Abstract

In order to personalize the behavior of hearing aid devices in different acoustic scenes, we need personalized acoustic scene classifiers. Since we cannot afford to burden an individual hearing aid user with the task to collect a large acoustic database, we will want to train an acoustic scene classifier on one in-situ recorded waveform (of a few seconds duration) per class. In this paper we develop a method that achieves high levels of classification accuracy from a single recording of an acoustic scene.

## 1 Introduction

We introduce our problem by an example. A hearing aid user may change several locations during the day, e.g., she may move through her home, car, office, grocery store, subway train etc. Since the acoustic conditions are quite different in each of these locations, the desired signal processing settings for her hearing aid may also differ across these locations. The set of frequently occurring acoustic environments is personal to a large extent. As a result, there is a need for personalized signal processing settings in a set of acoustic environments that differ across people.

In order to deal with this issue, we want a *personalizable* acoustic environment classifier as part of an intelligent hearing device. Since we want to impose as little burden on the end user as possible, we aim to build an acoustic environment classifier that can be trained under in-situ conditions by an end user who records a single example (of a few seconds in duration) of a new environment. We note that, aside from the hearing aids application, k-shot acoustic context* recognition can be also useful in other areas, such as urban monitoring and elderly care.

K-shot learning is relatively unexplored problem for acoustic scene classification. In this paper we develop a two-step procedure for training a probabilistic acoustic scene classifier by a single observed example of a novel acoustic environment. In the first step, the classifier is "pre-trained" on a broad acoustic database. This step sets the prior of the classifier to focus on acoustic scenes. In the second step, the classifier is further trained on the basis of a sample of a few seconds of a novel environment, which has been in-situ recorded by the end user.

We evaluate the performance of the proposed system on a benchmark dataset [1]. Our model features a competitive performance level with few training examples and beats a baseline method (Nearest-Neighbor classifier). Adding more training examples gradually improves the recognition accuracy.

## 2 Related Work

**k-shot learning** K-shot learning is a fairly well-studied problem in the context of computer vision ([2], [3], [4]) and more recently in the fields of generative modeling ([5], [6]) and reinforcement learning ([7]). Few works seem to relate to k-shot learning for acoustic modeling. [8] describes a

*We use the terms "environment", "context" and "scene" as synonyms in this paper.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

particularly relevant application to one-shot learning of speech concepts, where the trained model almost reaches human performance levels in recognizing new examples of words in both the Japanese and English languages. In [9], an approach to one-shot learning of arm gestures is described that is similar to the approach in the current paper.

**Acoustic scene modeling and classification** With the recent surge of interest in deep learning systems, many neural network-based models have been developed for acoustic scene classification, e.g., [10], [11], [12]. Other classification methods for this problem have also been investigated in the literature, however, they usually require hand-crafted feature extraction pipelines, which is something we want to avoid here. It is not clear yet whether modification of any of these approaches would result in a system that is able to outperform the proposed method in a one-shot learning mode.

## 3 Problem statement

Let $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,T})$ be an acoustic waveform that is drawn from an acoustic class $c \in C$. We assume to have access to a large set $X$ of *unlabeled* waveforms that were drawn from a wide range of scenes that effectively cover all classes in $C$. Later, we receive a new set of (in-situ obtained) *labeled* waveforms $D = \{(x_j, c_j)\}$, which were drawn from new scenes, i.e., $c_j \in C^+$, with $C^+ \cap C = \varnothing$. The data set $D$ contains $M \in \mathbb{N}^+$ waveforms for each scene $C^+$.

The task is to build a classifier that is able to correctly classify unseen waveforms from the new classes $C^+$, using information that is contained in $X$ and $D$. We will have a preference for very low values of $M$, e.g., $M = 1$ leads to the one-shot learning.

## 4 Model Specification

We use a generative probabilistic modeling approach, which requires, in our case, specification of a joint probability distribution $p(x, z, c, \theta|m)$ over the observed variables $x$, latent states $z$, latent classes $c$ and model parameters $\theta$ (and $m$ is a label for the model choice). All needed tasks, e.g., parameter estimation and classifier execution, can be formulated as inference tasks on this model. Omitting the conditioning on model $m$, we choose a model that factorizes as

$$p(x, z, c, \theta) = p(x, z|\theta, c)\, p(\theta|c)\, p(c)\,.$$

Natural signals consist of components that evolve over different time scales. This characteristic drives us toward hierarchical dynamic Bayesian models. We also want a mechanism to explicitly model the durations of staying in some meta-state (bird chirping, footsteps, etc.). Here, we use the Hidden Semi-Markov Model (HSMM) as the dynamics model since it appears to satisfy the basic requirements for dynamic modeling of natural acoustic sounds [13].

### 4.1 Hidden Semi-Markov Model

In an HSMM, a sequence $x$ is parsed into segments where a hidden segmental state remains constant over a variable number of time steps. Let $k \in \mathbb{N}^+$ be a segment counter. Each hidden segmental state $z_k \in \{1, 2, \ldots, S\}$ emits a variable number ($d_k$) of observations $x_{t_k}, x_{t_k+1}, \ldots, x_{t_k+d_k-1}$, where $d_k$ is drawn from a Poisson distribution. Since the j*th* segment contains $d_j$ samples, the first sample of the k*th* segment has time index $t_k = 1 + \sum_{j=1}^{k-1} d_j$. The dynamic part $p(x, d, z|\theta, c)$ of the generative model is formally described by[†]

$$\begin{aligned} p(x, d, z|\theta, c) &= p_c(x|z, d, \theta)\, p_c(d|z, \theta)\, p_c(z|\theta) \\ &= p_c(z_0) \prod_{k=1}^{K} \left( \prod_{t=t_k}^{t_k+d_k-1} \underbrace{\mathcal{N}\left(x_t \,\middle|\, \mu^{(c,z_k)}, \Sigma^{(c,z_k)}\right)}_{\text{observation}} \right) \underbrace{\text{Pois}\left(d_k \,\middle|\, \lambda^{(c,z_k)}\right)}_{\text{segment duration}} \underbrace{\text{Cat}\left(z_k \,\middle|\, \pi^{(c,z_{k-1})}\right)}_{\text{segment transition}} \end{aligned}$$

[†] *Contextual* information such as scene index $c$ and segmental index $z$ are collected in superscripts. *Temporal* indices such as time step $t$ and segment counter $k$ are denoted in subscripts.

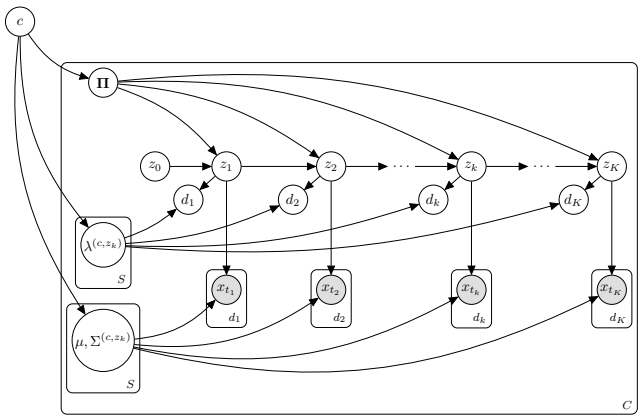


Figure 1: A Bayesian network graph of the model.

We specify the parameter priors by $\lambda^{(c,z_k)} \sim \text{Gam}\left(a^{(c,z_k)}, b^{(c,z_k)}\right)$, $\mu^{(c,z_k)} \sim \mathcal{N}\left(m^{(c,z_k)}, V^{(c,z_k)}\right)$, $\Sigma^{(c,z_k)} \sim \mathcal{W}^{-1}\left(\Psi^{(c,z_k)}\xi^{(c,z_k)}\right)$ and $\pi^{(c,z_{k-1})} \sim \text{Dir}\left(\phi^{(c)}\right)$ with $\phi^{(c)} \sim \text{Gam}(\alpha^{(c)}, \beta^{(c)})$.

Finally, we choose a uniform Categorical distribution for the prior over scene classes, making all acoustic scenes a priori equally likely, i.e., $p(c) = \text{Cat}\left(c \,\middle|\, \frac{1}{|C|}, \ldots, \frac{1}{|C|}\right)$. A graphical representation of this model is depicted in Figure 1.

## 5 Methods

**Learning** During learning, our goal is to infer the posterior distribution $p(\theta|c, X, D)$ for the model parameters for each class. Since the dataset $D$ contains very few ($M$) training examples per scene, we use a two-step procedure.

In the first step we learn a posterior distribution $p\left(\lambda^{(c,z_k)} \,\middle|\, c, X\right)$ for the segmental duration parameters using waveforms from $X$ by (Bayes rule):

$$p\left(\lambda^{(c,z_k)} \,\middle|\, c, X\right) \propto \sum_{z,d} \prod_{x_i \in X} p\left(x = x_i, d, z|c, \theta\right)\, p\left(\theta|c\right)\,. \tag{1}$$

In the second step we learn class-specific parameter distributions from labeled examples in $D$ by

$$p\left(\theta|c, X, D\right) \propto \sum_{z,d} \prod_{(x_j, c_j) \in D} p\left(x = x_j, d, z, c = c_j|\theta\right)\, p\left(\theta|c, X\right) \tag{2}$$

**Classification** We assign the class label $c^*$ to an unlabeled example $x^*$ based on the maximum posterior probability, i.e.,

$$c^* = \arg\max_{c \in C^+} p\left(c|x = x^*, X, D\right). \tag{3}$$

If we assume that all classes have same *a priori* probability $p(c)$, then the evaluation of $p\left(c|x = x^*, X, D\right)$ is equivalent to the evaluation of likelihood $p\left(x = x^*|c, X, D\right)$.

All inference and learning tasks were executed by Gibbs sampling as derived in [14].

## 6 Experimental evaluation

For the experimental evaluation we used the "TUT database for acoustic scene classification and sound event detection" (version 2016) that was collected by researchers at Tampere University of Technology [1]. The dataset contains 15 classes of acoustic scenes with 78 audio files of 30 seconds duration each. For each audio file, we calculated 20 Mel-Frequency Cepstral Coefficients (MFCC), plus delta and delta-delta derivatives (totaling 60 coefficients) for each window of 40 ms duration. For all HSMM models, the cardinality of the set of segmental states was set to $S = 20$.

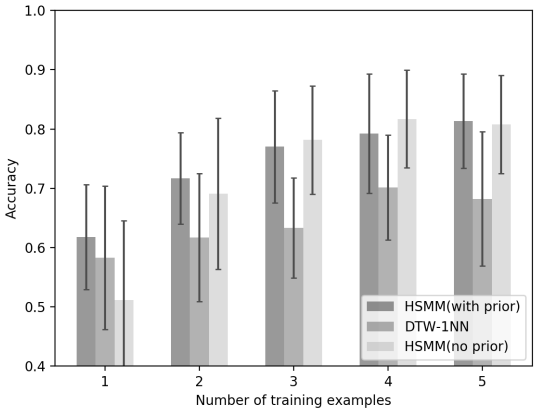


Figure 2: Classification accuracy as a function of number of training examples (1 standard deviation error bars over 20 repetitions).

The datasets were prepared by the following protocol:

1. We first randomly split the TUT dataset ([1]) into a subset `TUT-11` of 11 classes and the 4 remaining classes comprise `TUT-4`.
2. Next, we randomly drew one example from each of the 11 scenes in `TUT-11` to build an unlabeled data set $X$ that purposely covers a wide range of scenes.
3. We randomly drew $M$ examples (including labels) from each class in `TUT-4` to build the *labeled* training data set $D_{\text{train}}^{(M)}$. The remaining examples are grouped into $D_{\text{test}}^{(M)}$.

Training in step 1 on dataset $X$ and in step 2 on dataset $D_{\text{train}}^{(M)}$ proceeded by Eqs. 1 and 2 respectively. Classification performance on data set $D_{\text{test}}^{(M)}$ was evaluated by Eq. 3.

We repeated this process 20 times for each $M \in \{1, 2, 3, 4, 5\}$ to test how well our system performs on different combinations of selected classes. We also tested the performance of a baseline Nearest-Neighbor classifier with Dynamic Time-Warping (DTW) distance [15] and of an HSMM model without transfer of duration distributions (but still following the same protocol).

Our proposed method achieves almost 62% of classification accuracy in one-shot mode of learning, see Fig. 2. The Nearest-Neighbor classifier with DTW distance (which gets state-of-the-art results on time series classification tasks [16]) achieves slightly over 58%. We also note that it is beneficial to incorporate prior information in the learning process since the performance difference between the HSMM-with-prior and HSMM-without-prior is roughly 10% recognition accuracy.

## 7 Conclusions

K-shot learning of acoustic context is a challenging task that is important for scene-dependent personalization of hearing aid algorithms. In this paper, we report on a two-step procedure for data-efficient in-situ training of a Hidden Semi-Markov Model-based acoustic classifier. The model was evaluated on a real-world dataset of recorded acoustic scenes. The performance evaluation showed that incorporation of prior duration distributions is useful for the one-shot learning problem.

## Acknowledgments

This work is part of the research programme HearScan with project number 13925, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). This work also benefited from usage of the `pyhsmm` [17] package and we gratefully acknowledge the developers of that package.

## References


[1] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[2] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer vision and Image understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[3] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1842–1850.

[4] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 3630–3638.

[5] D. Rezende, S. Mohamed, I. Danihelka, K. Gregor, and D. Wierstra, "One-shot generalization in deep generative models," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1521–1529.

[6] R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba, "One-Shot Learning with a Hierarchical Nonparametric Bayesian Model." in *ICML Unsupervised and Transfer Learning*, 2012, pp. 195–206.

[7] Y. Duan, M. Andrychowicz, B. Stadie, J. Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba, "One-Shot Imitation Learning," *arXiv preprint arXiv:1703.07326*, 2017.

[8] B. M. Lake, C.-y. Lee, J. R. Glass, and J. B. Tenenbaum, "One-shot learning of generative speech concepts," in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 2014.

[9] A. van Diepen, M. Cox, and B. de Vries, "An In-situ Trainable Gesture Classifier," in *Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning*, Technische Universiteit Eindhoven, Jun. 2017, pp. 66–69.

[10] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for DCASE-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[11] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016.

[12] Y. Han and K. Lee, "Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[13] S.-Z. Yu, "Hidden semi-Markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243, Feb. 2010.

[14] M. J. Johnson and A. S. Willsky, "Bayesian nonparametric hidden semi-Markov models," *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 673–701, 2013.

[15] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[16] S. Seto, W. Zhang, and Y. Zhou, "Multivariate time series classification using dynamic time warping template selection for human activity recognition," in *2015 IEEE Symposium Series on Computational Intelligence*. IEEE, 2015, pp. 1399–1406.

[17] M. J. Johnson *et al.*, "pyhsmm," https://github.com/mattjj/pyhsmm, 2017.