Acoustic Scene Classification from Few Examples

Ivan Bocharov
Eindhoven University of Technology
Eindhoven, the Netherlands
Email: i.a.bocharov@tue.nl

Tjalling Tjalkens
Eindhoven University of Technology
Eindhoven, the Netherlands
Email: t.j.tjalkens@tue.nl

Bert de Vries
GN Hearing and TU Eindhoven
Eindhoven, the Netherlands
Email: bdevries@ieee.org

***Abstract*—In order to personalize the behavior of hearing aid devices in different acoustic environments, we need to develop 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 aim instead to train a scene classifier on just one (or maximally a few) in-situ recorded acoustic waveform of a few seconds duration per scene. In this paper we develop such a "one-shot" personalized scene classifier, based on a Hidden Semi-Markov model. The presented classifier consistently outperforms a more classical Dynamic-Time-Warping-Nearest-Neighbor classifier, and correctly classifies acoustic scenes about twice as well as a (random) chance classifier after training on just one recording of 10 seconds duration per scene.**

I. Introduction

The acoustic conditions around hearing aid (HA) users generally change multiple times throughout the day, e.g., HA users may move from and to their home, car, office, grocery store, subway train, etc. Oftentimes the preferred values for the HA tuning parameters depend on the acoustic context. Crucially, the set of acoustic scenes that drive preferred parameter settings differs across the HA user population. As a result, there is a need for a *personalizable* acoustic environment classifier in a modern hearing aid device. Since we want to inflict as little burden as possible on the end user, we aim to build an acoustic environment* classifier that can be trained under situated (in-situ) conditions by a HA user who only records a single (or maximally a few) example(s) of a few seconds in duration of any new environment.

Aside from the need to use very little training data, our application-in-mind (hearing aids and other wearables such as hearables) implies further constraints on the classifier. Firstly, in order to perform well on a recognition task for which only few labeled examples are available, the method needs to be based on strong inductive biases. These biases can come from several sources and in different forms. Notably, they can either be explicitly represented in model structure or they can come from a meta-learning procedure. Since our application is intended to execute in-situ on devices with small computational resources, we do not favor a recognition model that relies on computationally demanding meta-learning methods. Therefore, in this paper we focus on strong model assumptions that reflect our knowledge about properties of acoustic signals.

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

Another design consideration relates to the desire to share computational resources by multiple functional modules in low-power devices. For example, a noise power estimator inside a hearing aid algorithm may be used both for noise suppression and inside a scene classifier. This notion of sharing computational results sits well with a "generative probabilistic modeling" approach to classification. In this approach, all tasks (including scene classification) are formulated as probabilistic inference tasks on a generative model. This approach supports adding new inference tasks for alternative applications on the same generative model. As an additional advantage, the generative probabilistic modeling approach also facilitates a fast iterative design approach to the development of the scene classifier, since probabilistic inference is in principle an automatable task with a steadily improving toolset, e.g. [1] and [2].

In short, in this paper we develop a generative probabilistic model (based on a hidden Semi-Markov model (HSMM)) for personalizable acoustic scene classification that is suited to be executed under in-situ conditions on low-power devices. We train the HSMM classifier by a single observation of 10 seconds duration per acoustic environment. The HSMM model is equipped with strong biases in the form of priors and domain knowledge that are embedded in the model structure.

The performance of the proposed system was evaluated on a 2017 version of a benchmark dataset [3]. Our model consistently beats a baseline Nearest-Neighbor classifier with dynamic time warping alignment. Adding more training examples gradually improves the recognition accuracy.

II. Related Work

Learning from few examples is a relatively unexplored problem for acoustic scene classification. We shortly recapitulate some relevant studies.

a) Supervised learning from small datasets: Most existing studies concern the problem of learning from a small labeled dataset in context of k-shot learning. The k-shot learning problem statement assumes access to a big set of (possibly unlabeled) data. The data set is used to infer shared properties of exemplars of different categories in order to speed up learning from previously unseen object categories.

Different techniques for k-shot learning have been successfully employed for such tasks as image recognition [4], [5], generative modeling [6], and reinforcement learning [7].

ISBN 978-90-827970-1-5

Few works appear to relate learning from few examples to acoustic modeling. [8] describes a particularly relevant application to one-shot learning of speech concepts. Proposed hierarchical generative model trained on large Japanese and English speech corpora almost matches human performance in recognizing new words in both languages.

b) 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., [9], [10], [11]. These networks usually require a very large set of trainable parameters, which is not compatible with our objective to train on a very small data set. Moreover, the computational load of neural network-based classification often exceeds the resources of wearable devices. Alternative ("strong" model-based) classification methods for this problem have also been investigated, e.g. [12]. For our application it is important to keep feature extraction pipeline as simple as possible, which is usually not the case for these methods. Our model shares the design philosophy of the strongly-biased generative probabilistic acoustic model proposed by Lee and Glass [13], but the model specifications details are different (as well as the application).

III. Problem statement

Let $\mathcal{X}$ be a set of sequences and $\mathcal{C}$ a finite set of class labels. A sequence x_i is an ordered tuple of observations: $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,T}) \in \mathcal{X}$. We assume to have access to a *training set* of (in-situ obtained) *labeled* sequences $\mathcal{D} = \{(x_j, c_j)\}$, where $x_j \in \mathcal{X}$ and $c_j \in \mathcal{C}$. The data set $\mathcal{D}$ contains $M \in \mathbb{N}^+$ sequences for each class $c \in \mathcal{C}$.

The task is to build a classifier $f : \mathcal{X} \to \mathcal{C}$ that is able to correctly classify unseen sequences from classes $\mathcal{C}$, using information that is contained in $\mathcal{D}$. Since in our application the dataset has to be obtained under in-situ conditions, we will have a preference for very low values of M.

IV. Model Specification

In this paper we use a generative probabilistic modeling approach, which requires 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 θ (and m is a label for the model choice). After the model has been specified, all needed tasks, e.g., parameter estimation and classifier execution, can then be formulated as inference tasks on this model. Omitting the conditioning on model m, we choose a model that factorizes as

$$p(x, z, c, \theta) = \underbrace{p(x, z|\theta, c)}_{\text{dynamics}} \cdot \underbrace{p(\theta|c)}_{\text{parameters}} \cdot \underbrace{p(c)}_{\text{scene prior}} .$$

Our choice of dynamics model is motivated by the hierarchical nature of real-world audio signals. An acoustic scene can be represented as a sequence of changing meta-states (clatter of plates, keyboard clacking, etc.). At the same time, the duration of staying in the same state might be different for different states. For this reason, having a mechanism that explicitly models these durations might be useful. In this paper, we use the Hidden Semi-Markov Model (HSMM) [14] as the dynamics model since it appears to satisfy our requirements for dynamic modeling of natural acoustic sounds. For instance, in an HSMM dynamic acoustic model the signal components evolve over multiple timescales: on the order of milliseconds at the (bottom) observation level, up to hundreds of milliseconds at the (middle) hidden segmental state level to seconds and minutes at the (top) scene level.

A. 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 state-specific 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

$$\begin{aligned} p(x, d, z|\theta, c) &= p_c(x|z, d, \theta)\, p_c(d|z, \theta)\, p_c(z|\theta) \\ &= p_c(z_0) \prod_{k=1}^{K} \left(\prod_{t=t_k}^{t_k+d_k-1} \underbrace{p_c\left(x_t \mid z_k, \theta\right)}_{\text{observation}} \right) \\ &\quad \cdot \underbrace{p_c\left(d_k \mid z_k, \theta\right)}_{\text{segment duration}} \underbrace{p_c\left(z_k \mid z_{k-1}, \theta\right)}_{\text{segment transition}} . \end{aligned}$$

The Hidden Semi-Markov model consists of three key parts: an observation model $p_c\left(x_t \mid z_k, \theta\right)$, a segment duration model $p_c\left(d_k \mid z_k, \theta\right)$ and a segment transition model $p_c\left(z_k \mid z_{k-1}, \theta\right)$. The flexibility provided by the model allows us to choose the distributions according to our assumptions. We use a Poisson distribution to model the durations, a categorical state transition model (similar to regular HMMs) and a full-rank multivariate normal distribution for observations:

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

In this model description, *contextual* information such as scene index c and segmental index z is collected in superscripts. *Temporal* indices such as time step t and segment counter k are denoted in subscripts. For instance, $\mathrm{Cat}\left(z_k \,\middle|\, \pi^{(c,z_{k-1})}\right)$ indicates a categorical distribution where the entry at index $(k-1, k)$ in the transition matrix specifies $p(z_k|z_{k-1}) = \pi^{(c,z_{k-1})}$. We assume that the diagonal elements of the transition matrix are equal to zero in order to avoid self-transitions between states. If we allow self-transitions, duration distributions do not model state duration statistics directly.

ISBN 978-90-827970-1-5

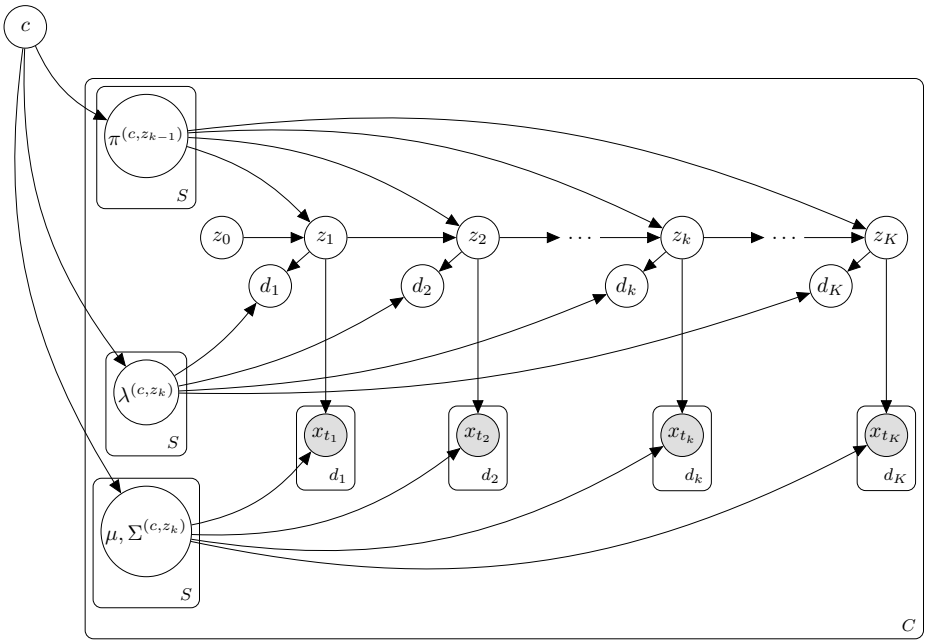

Fig. 1. A Bayesian network graph of the HSMM-based acoustic classifier that is discussed in this paper.

We specify the parameter priors by

$$\begin{aligned}
\lambda^{(c,z_k)} &\sim \mathrm{Gam}\left(a^{(c,z_k)}, b^{(c,z_k)}\right), \\
\mu^{(c,z_k)} &\sim \mathcal{N}\left(m^{(c,z_k)}, V^{(c,z_k)}\right), \\
\Sigma^{(c,z_k)} &\sim \mathcal{W}^{-1}\left(\Psi^{(c,z_k)}, \xi^{(c,z_k)}\right), \\
\pi^{(c,z_{k-1})} &\sim \mathrm{Dir}\left(\phi^{(c)}\right), \\
\text{with } \phi^{(c)} &\sim \mathrm{Gam}\left(\alpha^{(c)}, \beta^{(c)}\right).
\end{aligned}$$

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

B. Dynamic Time Warping

We will use a dynamic time warping-based model as a reference acoustic context learning model. Dynamic time warping (DTW) [15] is a dynamic programming-based method that aligns two sequences by warping the temporal dimension. DTW alignment is usually used as a preprocessing step for calculating a distance measure between two sequences if we are not interested in dissimilarities that are caused by temporal warping of the sequences. For example, this method may be used to measure the similarity ("distance") between two persons in how they ride a bicycle, regardless of the bikes' speeds and possible accelerations/decelerations. The Nearest-Neighbor (NN) method in conjunction with DTW alignment generally performs well on time series classification tasks, as reported in [16]. This qualifies the DTW-NN method to be considered a worthy adversary in the task of acoustic context classification from a small training data set.

V. Methods

In generative probabilistic models, all tasks are formulated as inference tasks. We use Gibbs sampling derived in [17] to perform inference for classification and learning.

a) Learning: During learning, our goal is to infer the posterior distribution $p(\theta|c, \mathcal{D})$ for the model parameters for each class. We learn class-specific parameter distributions from labeled examples in D by

$$p\left(\theta|c, \mathcal{D}\right) \propto \sum_{z,d} \prod_{(x_j, c_j) \in D} p\left(x = x_j, d, z, c = c_j|\theta\right) p\left(\theta|c\right). \quad (1)$$

b) Classification: The Bayesian solution points to inferring the posterior class probability $p(c|x = x^*, \mathcal{D})$ for a given sequence x^*. We then assign the class label c^* by maximizing the posterior probability:

$$c^* = \arg\max_{c \in \mathcal{C}} p\left(c|x = x^*, \mathcal{D}\right). \quad (2)$$

In our case the evaluation of $p\left(c|x = x^*, \mathcal{D}\right)$ is equivalent to the evaluation of the likelihood $p\left(x = x^*|c, \mathcal{D}\right)$, since all classes have same a priori probability $p(c)$.

VI. Experimental evaluation

A. Dataset preparation

For the experimental evaluation we used the "TUT database for acoustic scene classification and sound event detection" (version 2017) that was collected by researchers at Tampere

ISBN 978-90-827970-1-5

TABLE I
CLASSIFICATION ACCURACY

	Number of training examples				
	1	2	3	4	5
HSMM	0.514 ± 0.137	0.597 ± 0.089	0.636 ± 0.084	0.666 ± 0.069	0.718 ± 0.080
DTW-NN	0.443 ± 0.132	0.470 ± 0.101	0.489 ± 0.088	0.551 ± 0.081	0.522 ± 0.092

University of Technology [3]. The TUT database contains recordings of 15 different acoustic scenes. The scenes in the database cover a wide variety of real-world environments and can be split into the meta-categories "inside", "outside" and "inside a vehicle".

The TUT database contains a development database of 312 audio files (10 seconds duration each) for each acoustic scene as well as an evaluation dataset of 108 recordings (also 10 seconds each) for each scene.

In order to prepare training and evaluation datasets, we first randomly selected 4 scenes from the TUT database, comprising the set $\mathcal{C}$. The training dataset D_{train} was formed by randomly drawing M exemplars (waveforms plus scene labels) from the TUT development database for each acoustic scene in $\mathcal{C}$. In order to test the performance of the classifier, an evaluation set $\mathcal{X}_{\text{eval}}$ was formed by collecting all examples from scenes $\mathcal{C}$ from the TUT evaluation database.

B. Data preprocessing

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 with 20 ms hop length. These coefficients aggregate the behavior of the signal at the order of tens of milliseconds, thus extending the hierarchy in the model with an additional layer. Since the map from waveform samples to MFCC coefficients is deterministic, we used the MFCC sequences (rather than the raw waveforms) as observations for the (HSMM and DTW-NN) classifiers.

C. Model priors and hyperparameters

TABLE II
PARAMETER PRIORS FOR EXPERIMENTAL EVALUATION

Parameter	Prior distribution
$\lambda^{(c,z_k)}$	$\text{Gam}\,(60, 2)$
$\mu^{(c,z_k)}$	$\mathcal{N}\,(0, 100 * I)$
$\Sigma^{(c,z_k)}$	$\mathcal{W}^{-1}\,(0.25, 62)$
$\pi^{(c,z_{k-1})}$	$\text{Dir}\,(\phi^{(c)})$
$\phi^{(c)}$	$\text{Gam}(1.0, 0.25)$

For all HSMM-based classifiers, the cardinality of the set of segmental states was set to $S = 20$. See Table II for a detailed list of experiment-specific prior settings.

D. Evaluation protocol

The HSMM classifier was trained on dataset $\mathcal{D}_{\text{train}}$ by executing Eq. 1. Performance assessment was executed by classifying the evaluation dataset $\mathcal{X}_{\text{eval}}$ by Eq. 2. We used the fraction of correctly classified scenes as score function,

$$accuracy(c^*, c) = \frac{1}{|X_{eval}|} \sum_{i=1}^{|X_{eval}|} I(c_i^* = c_i),$$

where $I(s) = 1$ if $s = \texttt{true}$, and otherwise $I(s) = 0$.

We evaluated the classifier for each $M \in \{1, 2, 3, 4, 5\}$. In order to get evaluation results that do not depend on a particular draw of scenes from the TUT database, we repeated the entire evaluation process 20 times and report both the mean value and standard deviation of the classifier score. The evaluation of the DTW-NN classifier followed the same procedure. Algorithm 1 describes our evaluation protocol in pseudo-code.

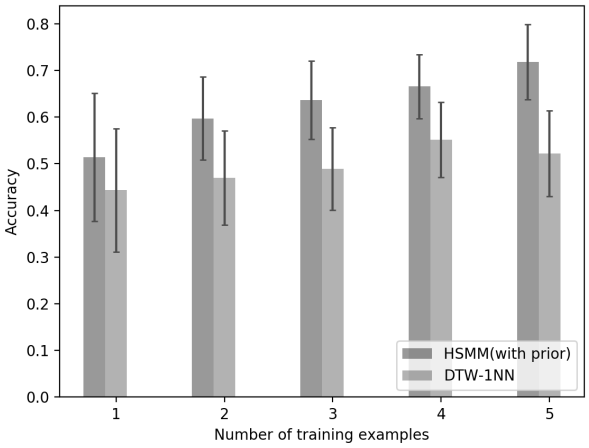

Fig. 2. Classification accuracy as a function of number of training examples (± 1 standard deviation error bars over 20 repetitions).

E. Evaluation results

The HSMM classifier achieves 51% of classification accuracy in a one-shot learning mode, see Fig. 2 and Table I. The Nearest-Neighbor classifier with DTW distance achieves slightly over 44%. Adding more training examples gradually improves the performance of both classifiers, while the HSMM classifier remains clearly preferable over the DTW-NN classifier. When presented with 5 labeled recordings for each class (20 recordings total), the developed HSMM classifier has reached an accuracy score of 0.71.

ISBN 978-90-827970-1-5

Algorithm 1 Evaluation protocol

Require: Datasets `TUT-DEV`, `TUT-EVAL`
  for all $M \in \{1, 2, 3, 4, 5\}$ **do**
    for n=1..20 **do**
      $\mathcal{C} \coloneqq$ sample 4 categories from `TUT-DEV`
      $D_{train} \coloneqq$ sample M exemplars from each class in $\mathcal{C}$
      $X_{eval}, c_{true} \coloneqq$ all examples of classes $\mathcal{C}$ from `TUT-EVAL`
      for all classifiers $f \in \{\texttt{HSMM}, \texttt{DTWNN}\}$ **do**
        train classifier f on D_{train}
        $c_{pred} \coloneqq$ evaluate classifier f on X_{eval}
        $scores[M][n][f] \coloneqq accuracy(c_{pred}, c_{true})$
      end for
    end for
  end for
  return $scores$

VII. Discussion

The proposed probabilistic classifier for acoustic scenes correctly recognizes the acoustic environment from a single training example in around a half of the cases (see Table I). This is about twice as good as a (random) chance classifier.

The classifier performance scores are affected by the fact that the evaluation datasets often contained similar categories. For instance, to successfully distinguish between a library and an office scene from a single recording of 10 seconds is a very difficult problem. At the same time, an error of this kind might not result in worse performance in an application environment since a hearing aid user may prefer the same settings for audio signal processing in both library and office environments. We consider the hypothesis that the currently described in-situ personalizable HSMM-based classifier already leads to improved hearing aid user satisfaction scores as a topic for further investigation.

VIII. Conclusions

In-situ learning of an acoustic classifier from single (or few) short recordings of acoustic waveforms is a very challenging task. In this paper we presented a generative probabilistic modeling approach, specifically based on a Hidden Semi-Markov Model, to the design of a scene classifier that can be trained on very few recordings. We showed that the proposed HSMM classifier consistently beats a reference DTW-NN classifier and scores about twice as well as random classification after training on a single example per scene.

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). We also gratefully acknowledge the developers of the `pyhsmm` [18] package, which was helpful to execute the inference tasks.

References

[1] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, vol. 76, no. 1, 2017.

[2] D. Tran, M. D. Hoffman, R. A. Saurous, E. Brevdo, K. Murphy, and D. M. Blei, "Deep Probabilistic Programming," in *International Conference on Learning Representations*, Mar. 2017.

[3] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[4] 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.

[5] 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.

[6] 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.

[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] 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.

[10] 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.

[11] 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.

[12] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2013, pp. 1–4.

[13] C.-y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

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

[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] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures," *Proc. VLDB Endow.*, vol. 1, no. 2, pp. 1542–1552, Aug. 2008.

[17] 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.

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