RECURRENT NEURAL NETWORKS WITH STOCHASTIC LAYERS FOR ACOUSTIC NOVELTY DETECTION

Duong Nguyen [1] *, Oliver S. Kirsebom* [2]*, Fábio Frazão* [2]*, Ronan Fablet* [1]*, Stan Matwin* [2,3]

(1) IMT Atlantique, Lab-STICC, UBL, Brest, France
(2) Institute for Big Data Analytics, Dalhousie University, Halifax, Canada
(3) Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

ABSTRACT

In this paper, we adapt Recurrent Neural Networks with Stochastic Layers, which are the state-of-the-art for generating text, music and speech, to the problem of acoustic novelty detection. By integrating uncertainty into the hidden states, this type of network is able to learn the distribution of complex sequences. Because the learned distribution can be calculated explicitly in terms of probability, we can evaluate how likely an observation is then detect low-probability events as novel. The model is robust, highly unsupervised, end-to-end and requires minimum preprocessing, feature engineering or hyperparameter tuning. An experiment on a benchmark dataset shows that our model outperforms the state-of-the-art acoustic novelty detectors.

Index Terms— acoustic modeling, novelty detection, variational recurrent neural network, stochastic recurrent neural network.

1. INTRODUCTION

Audio processing in general, and acoustic novelty detection in particular has attracted significant attention recently. A number of studies have used acoustic data to detect abnormal events, mostly for surveillance purposes, such as human fall detection [1], [2], abnormal jet engine vibration detection [3], hazardous events detection [4].

The main challenge of novelty detection is we do not have a large amount of novel events to learn their characteristics, while the normal set is usually very big and contains a large amount of uncertainty. The common approach is to use unsupervised methods to learn the normality model, then consider events that do not fit this model as abnormal (novel). Most of these systems use Gaussian Mixture Model (GMM) or Hidden Markov Model (HMM) [5], [4], [6]. Bayesian Networks have also been explored [7], [8]. Recently, advances in deep learning [9], especially in Recurrent Neural Networks (RNNs) and their extensions (Long Short-Term Memory — LSTM [10], Gated Recurrent Unit — GRU [11]) have opened new venues for acoustic modeling. In [12], the authors employed LSTMs to create an AutoEncoder (AE) to model normal sounds and detect abnormal sounds using the reconstruction errors. This idea has been extended in [13] by applying an adversarial training protocol.

However, acoustic signals are stochastic. RNN-based networks, whose hidden states are deterministic, can hardly capture all the variations in the data. Recent efforts to improve the modeling capacity of RNNs by including stochastic factors in their hidden states have shown impressive results, especially for generating text, music and speech [14], [15], [16], [17].

In this paper, we adapt these models to create an unsupervised acoustic novelty detector. Our approach performs an end-to-end learning of a probabilistic representation of acoustic signals. Given this representation, we can evaluate how likely an observation and state the detection of novel events as the detection of observations with a low probability. We argue that this model is robust, highly unsupervised, end-to-end and requires minimum preprocessing, feature engineering or hyperparameter tuning. Our empirical evaluation on a dataset for novel event detection in audio data shows that the proposed model outperforms the state-of-the-art.

The paper is organized as follows: in Section 2, we present the details of the proposed approach; we compare the model with state-of-the-art methods to point out its advantages in Section 3; the experiment and results are shown in Section 4; finally in Section 5 we give conclusions and some perspectives for future work.

2. THE PROPOSED APPROACH

2.1. Recurrent Neural Networks with Stochastic Layers (RNNSLs)

For time series modeling, the two most common approaches are State Space Models (SSMs) and Recurrent Neural Networks (RNNs). SSMs such as Kalman filters [18] and particle filters [19] have been explored for a long time and are the

This work was supported by the UBL Mobility Fund, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Labex Cominlabs, the Brittany Council and the "Groupement Bretagne Télédétection (BreTel)

The authors would like to thank A3Lab for the dataset used in this paper.

state-of-the-art model-driven schemes thanks to their ability to model stochasticity. However, these models are limited by their mathematical assumptions (for example, Kalman filters assume the data generating process is Gaussian). RNNs, on the other hand, have attracted a lot of attentions recently by their capacity to represent long-term dependencies in time series [9]. The main drawback of RNNs is that their hidden states are deterministic, making them unable to capture all the stochastic components of the data. A number of efforts have been made to bring together the power of SSMs and RNNs [14], [16], [17], [20]: Recurrent Neural Networks with Stochastic Layers (RNNSLs).

RNNSLs aim to learn the distribution p, which can be factored through time, over a sequence of T observed random variables $\{\mathbf{x}_t\}_{,t=1..T}$:

$$p(\mathbf{x}_{1:T}) = \prod_{t=1}^{T} p_t(\mathbf{x}_t|\mathbf{x}_{<t}), \quad (1)$$

where $\mathbf{x}_{<t}$ denotes $\mathbf{x}_{1:t-1}$.

Following a SSM formulation, we assume that the data generation process of $\mathbf{x}_{1:T}$ relies on a sequence of T latent random variables $\{\mathbf{z}_t\}_{,t=1..T}$. At each time step t, the joint distribution $p_t(\mathbf{x}_t, \mathbf{z}_t|\mathbf{x}_{<t}\mathbf{z}_{<t})$ can be factored into:

$$p_t(\mathbf{x}_t, \mathbf{z}_t|\mathbf{x}_{<t}\mathbf{z}_{<t}) = p_t(\mathbf{x}_t|\mathbf{x}_{<t}, \mathbf{z}_{\leq t})p_t(\mathbf{z}_t|\mathbf{x}_{<t}, \mathbf{z}_{<t}), \quad (2)$$

where $\mathbf{z}_{\leq t}$ denotes $\mathbf{z}_{1:t}$. In other words, each time step of the network is an autoencoder, conditionally to the historical information.

Depending on the stochastic nature of the considered data, the emission distribution $p_t(\mathbf{x}_t|\mathbf{x}_{<t}, \mathbf{z}_{\leq t})$ may be highly nonlinear. However, this nonlinearity usually leads to the intractability of the inference distribution $p_t(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t})$. The most common solution to overcome this obstacle is the variational approach [15], [16], [17], which introduces an approximation $q_t(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t})$ of the posterior distribution $p_t(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t})$ then estimates $p_t(\mathbf{x}_t|\mathbf{x}_{<t})$ by the Evidence Lower BOund (ELBO) $\mathcal{L}(\mathrm{x}, p_t, q_t)$:

$$\begin{aligned}\log p_t(\mathbf{x}_t|\mathbf{x}_{<t}) \geq \mathcal{L}(\mathrm{x}, p_t, q_t) = \\ \mathbb{E}_{\mathbf{z}_t\sim q_t}\big[\log p_t(\mathbf{x}_t|\mathbf{x}_{<t}, \mathbf{z}_{\leq t})\big] \\ - \mathrm{KL}\big[q_t(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t})||p_t(\mathbf{z}_t|\mathbf{x}_{<t}, \mathbf{z}_{<t})\big] \quad (3)\end{aligned}$$

where $\mathrm{KL}\big[q_t||p_t\big]$ is the Kullback-Leibler divergence between two distributions q_t and p_t.

There are several types of RNNSLs, differing in the way that they model the structure of the latent space. The most common types are Variational Recurrent Neural Networks (VRNNs) [16], Stochastic Recurrent Neural Networks (SRNNs) [17] and Deep Kalman Filters (DKFs) [20]. We experimented most of these types, however, in this paper, for simplicity purposes, we only report the VRNNs, introduced by Chung et al. [16].

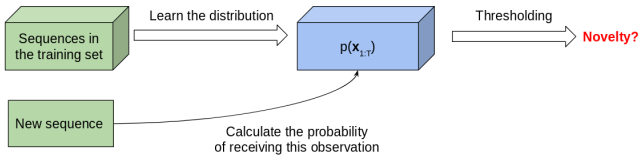

Fig. 1. Architecture of the proposed RNNSL-based novelty detector.

In VRNNs, the historical information $(\mathbf{x}_{<t}, \mathbf{z}_{<t})$ is encoded by the dynamics of the hidden states of their RNN (LSTM) $\mathbf{h}_t = h(\mathbf{x}_{t-1}, \mathbf{z}_{t-1}, \mathbf{h}_{t-1})$. More precisely, it involves the parameterization of the following distributions, namely the emission distribution $p_t(\mathbf{x}_t|\mathbf{x}_{<t}, \mathbf{z}_{\leq t}) = p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_t)$, the prior distribution $p_t(\mathbf{z}_t|\mathbf{x}_{<t}, \mathbf{z}_{<t}) = p(\mathbf{z}_t|\mathbf{h}_t)$ and the variational posterior distribution $q_t(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t}) = p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_t)$ as neural networks. Here, we consider fully connected networks with Gaussian formulation of these three distributions. For more details of VRNNs, please refer to [16].

2.2. RNNSLs for Acoustic Novelty Detection

RNNSLs were initially designed for generating text, music, speech. They are currently the state-of-the-art in these domains [15], [16], [17], [21]. The interesting point of this type of models in comparison to other state-of-the-art methods like Wavenet [22] is that these models calculate the distribution $p(\mathbf{x}_{1:T})$ explicitly, so that after learning this distribution from the training set, we can evaluate the probability for each new sequence. The idea of using RNNSLs for novelty detection was first introduced in [23] for the detection of abnormal behaviors of vessels, we adapt this model to novelty detection in acoustic data.

Here, an acoustic signal is modeled as a time series $\{\mathbf{x}_t\}_{,t=1..T}$ where $\mathbf{x}_t$ can be a chunk of n samples of the waveform, or n frequency bins in a spectrogram at a given time t. A RNNSL first learns the distribution over $\mathbf{x}_{1:T}$ in the training set, which may or may not contain some abnormal sequences. Then, for any new acoustic signal, we can evaluate its log-probability. If this log-probability is smaller than a threshold, the sequence will be considered as abnormal (or novel), as illustrated in Fig. 1.

To choose the threshold, we create a validation set, which again may or may not contain some abnormal sequences and compute the mean μ_{valid} and the standard deviation σ_{valid} of the log-probability of the sequences in this set. The value of the threshold is then chosen as: $\theta = \mu_{valid} - \alpha * \sigma_{valid}$. α is usually chosen as 3.

The training set and the validation set may contain some abnormal sequences. However, since RNNSLs are probabilistic models, they will eventually ignore these "outliers" (this conjecture is confirmed experimentally). This property helps to reduce data cleaning efforts.

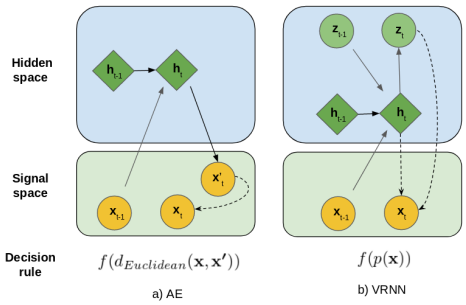

Fig. 2. Architecture and decision rule of the proposed model (VRNN) in compared to previously proposed AE-based models. $\mathbf{x}_t$ is the original signal at the given time step t, $\mathbf{h}_t$ is the hidden state of the RNN (LTSM), $\mathbf{z}_t$ is the latent stochastic state, $\mathbf{x}'_t$ is the reconstructed output of the AE. The solid arrows denote the calculation processes, while the dashed arrows show how the cost function is calculated. We use the same notation as [17], circles for stochastic factors, diamonds for deterministic factors.

3. RELATED WORK

A number of researches have explored deep neural networks to detect novelty in acoustic surveillance. We point out here the advantages of our model over those used in [12] and [13], which are currently the state-of-the-art methods.

Both [12] and [13] used RNNs (LSTMs in particular) as an AutoEncoder (AE) which can reconstruct the original signal from a compressed representation (Compression AutoEncoders — CAEs) or from a corrupted version of it (Denoising AutoEncoders — DAEs). However, as discussed in [16], [17] and [20], the fact that the hidden states of RNNs are deterministic reduces their capacity to capture all data variabilities, especially for data that contain high levels of randomness.

Moreover, the detection criterion used in [12] is the Euclidean distance between the original input and the reconstructed output of the autoencoder. This criterion is very sensitive to noise. [13] addressed this drawback by using an adversarial strategy, however, the ultimate idea is also to compare the original input and the reconstructed output from the autoencoder. By contrast, our method detects novel events by directly evaluating the probability of the received signal. Besides the improved detection criterion, the architecture of our model is also more robust to noise [23].

These differences are sketched in Fig. 2. The hidden space of our model has stochastic factors, which help to increase modeling capacity. The decision rule of our model is a function of the distribution learned by the network, making the model more robust to noise.

The selection of the thresholding value for novelty detection is another important difference compared to previous works. The approach in [12] is not fully unsupervised, because it needs some information about the proportion of abnormal events in the data. Our method, in contrast, only uses the information from the training set and the validation set to chose the threshold, without any prior knowledge of the annotations, based on a statistically-sound criterion, *i.e.* the false alarm rate.

4. EXPERIMENT AND RESULT

4.1. Dataset

We tested our model[1] on the same dataset used in [12] and [13], which is part of the PASCAL CHiME speech separation and recognition challenge dataset [24]. The original dataset contains 7 hours of in-home environment recordings with two children and two adults performing common activities, such as talking, eating, playing and watching television. The author of [12] took a part of those recordings and created a dataset for acoustic novelty detection (100 minutes for the training set and 70 minutes for the test set). In the new dataset, the sounds of the PASCAL CHiME are considered as background, the test set was generated by digitally adding abnormal sounds like alarms, falls, fractures (breakages of objects), screams. The details of the dataset were presented in [12].

4.2. Experimental Setup

In order to use the models in [12] and [13] as baselines, we set up our model to have the same evaluation metric that was used in those papers. However, instead of transforming the data to mel spectrograms like in [12] and [13], we worked directly with the waveform (end-to-end model). The dataset was recorded by a binaural microphone at a sample rate of 16kHz. We converted each audio to 1 channel and then split it into sequences of 160-dimensional frames, each frame corresponds to 0.01s, as in [12] and [13]. [12] and [13] evaluated the detection at each frame instead of at the whole sequence, so we also applied the thresholding step to each $\log p(\mathbf{x}_t|\mathbf{x}_{<t})$, instead of $\log p(\mathbf{x}_{1:T})$.

We tested different topologies of VRNN, with the latent size of 64, 80, 160 and 200. The models were trained using Adam optimizer [25], with a learning rate of $3e-5$.

4.3. Results

Different configurations gave different log-likelihoods on the dataset, however the final detection results were quite similar. We report here only one of the topologies, which gave the best result: VRNN with 160 latent units (the models with 80 hidden units also gave similar results). We compare the performance of our model with the result of GMM, HMM, those in [12] (LSTM-based CAE, LSTM-based DAE) and in [13] (Adversarial AE). The result is shown in Table 1[2]. Besides choosing the threshold automatically as discussed in Section

[1] The code is available at https://github.com/dnguyengithub/AudioNovelty

[2] The values in Table 1 are from [12] and [13], [13] did not show the precision and recall of their model

Table 1. Detection result, in comparison with state-of-the-art methods.

Method	Online Processing	Precision	Recall	F1 score
GMM	Yes	99.1	87.8	89.4
HMM	Yes	94.1	88.9	91.1
LSTM-CAE	Yes	91.7	86.6	89.1
BLSTM-CAE	No	93.6	89.2	91.3
LSTM-DAE	Yes	94.2	90.6	92.4
BLSTM-DAE	No	94.7	92.0	93.4
Adversarial AE	?	?	?	93.3
VRNN	Yes	95.4	91.8	**93.6**
VRNN*	Yes	95.4	92.8	**94.1**

Table 2. Robustness test.

SNR	Precision	Recall	F1 score
5dB	96.0	91.2	93.6
10dB	96.1	91.9	94.0
15dB	96.1	92.1	94.0

2, we also used the same technique as in [12] to chose the optimal threshold value, denoted as **VRNN*** .

Our method not only outperformed the state-of-the-art methods, but also has the ability to work online, which is highly beneficial for real-time surveillance. Models that use bidirectional LSTM (BLSTM-CAEs, BLSTM-DAEs) can not reach online processing the because a look-ahead buffer is required. The online processing ability of Adversarial AEs depends on the structure that they use (LSTM or BLSTM).

When investigating the cases where the proposed model misdetected the novelty, we found that actually the model could detect all the novel events, however, the way the detection was evaluated reduced the accuracy. As in [12] and [13], the detection was evaluated at each time step of 0.01s. Our model has a memory effect (the memory of its LSTM cells), so it tends to merge the abnormal events that are very close to each other, as shown in Fig. 3. In other cases, the model missed a part of the sound, especially for the tail of the fractures, as shown in Fig. 4. These sounds have a long tail which is gradually submerged in the background. These misdetections are not detrimental in real life applications, because we are more interested in whether or not there is a novel event than on how long the event is.

We also conducted a robustness test where we added Gaussian noise to the test set. The additive noise is unknown by the model. This is a common scenario in audio surveillance, when the background environment changes (*e.g.* because of winds) or when noise appears in the electronic system. Table 2[3] shows the performance of the proposed approach (with optimal threshold) on the corrupted test sets with different level of Signal to Noise Ratio (SNR). Thanks to the nature of VRNNs and the improved detection criterion, our model is robust to noise.

[3] [12] and [13] did not provide sufficient detail to replicate their codes for this test.

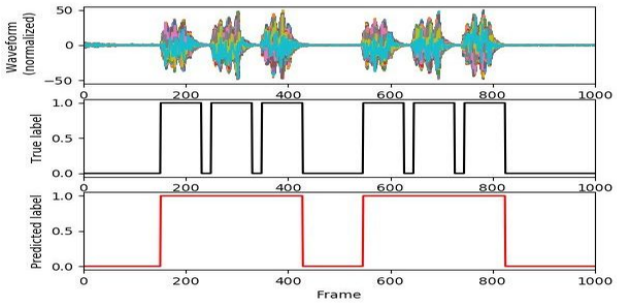

Fig. 3. An example where the novelty events were merged. This figure shows the waveform of two alarms, each alarm consists of there "beeps", our model considered this "beep beep beep" as one event, while the annotation made by the authors of [12] separates these "beeps".

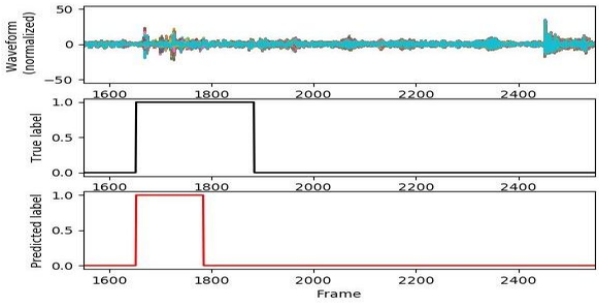

Fig. 4. An example where the model missed a part of the novelty event. This figure shows the waveform of the sound of a fracture of a dish. The tail of the sound is very mall and gradually becomes submerged in the background.

5. CONCLUSIONS AND PERSPECTIVES

We have presented a novel unsupervised end-to-end approach for acoustic novelty detection. This approach exploits RNNs with stochastic layers, which are the state-of-the-art frameworks for time series modeling. Given the learned probabilistic representations, novelty detection can be stated as a classic statistical test, which fully accounts for the stochasticity of the considered acoustic datasets. Reported experiments on a benchmarked dataset showed that the model outperforms the state-of-the-art detectors [12], [13].

The dataset used in this paper is quite simple, the novel events in it are quite easy to be detected. Future work could involve applying this model to more complex signals, *e.g.* underwater acoustic signals which depict even greater variabilities. The impact of the threshold is also being studied to obtain better threshold selection rule.

6. REFERENCES

[1] X. Zhuang, J. Huang, G. Potamianos, and M. Hasegawa-Johnson, "Acoustic fall detection using Gaussian mixture models and GMM supervectors," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Taipei, Taiwan: IEEE, Apr. 2009, pp. 69–72.

[2] M. Salman Khan, M. Yu, P. Feng, L. Wang, and J. Chambers, "An unsupervised acoustic fall detection system using source separation for sound interference suppression," *Signal Processing*, vol. 110, pp. 199–210, May 2015.

[3] D. A. Clifton and L. Tarassenko, "Novelty detection in jet engine vibration spectra," *International Journal of Condition Monitoring*, vol. 5, pp. 2–7, Aug. 2015.

[4] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Probabilistic Novelty Detection for Acoustic Surveillance Under Real-World Conditions," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 713–719, Aug. 2011.

[5] P. Kumar, A. Mittal, and P. Kumar, "A Multimodal Audio Visible and Infrared Surveillance System (MAVISS)," in *2005 3rd International Conference on Intelligent Sensing and Information Processing*, Dec. 2005, pp. 151–156.

[6] P. Atrey, N. Maddage, and M. Kankanhalli, "Audio Based Event Detection for Multimedia Surveillance," in *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, vol. 5. Toulouse, France: IEEE, 2006, pp. V–813–V–816.

[7] W. Zajdel, J. Krijnders, T. Andringa, and D. Gavrila, "CASSANDRA: audio-video sensor fusion for aggression detection," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. London, UK: IEEE, Sep. 2007, pp. 200–205.

[8] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-Visual Fusion for Detecting Violent Scenes in Videos," in *Artificial Intelligence: Theories, Models and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, vol. 6040, pp. 91–100.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[10] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," in *Fifteenth annual conference of the international speech communication association*, 2014, p. 5.

[11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated Feedback Recurrent Neural Networks," in *International Conference on Machine Learning*, 2015, p. 9.

[12] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 1996–2000.

[13] E. Principi, F. Vesperini, S. Squartini, and F. Piazza, "Acoustic novelty detection with adversarial autoencoders," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 3324–3330.

[14] J. Bayer and C. Osendorfer, "Learning Stochastic Recurrent Networks," *arXiv:1411.7610 [cs, stat]*, Nov. 2014, arXiv: 1411.7610.

[15] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription," in *29th International Conference on Machine Learning (ICML 2012)*, Jun. 2012, arXiv: 1206.6392.

[16] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio, "A Recurrent Latent Variable Model for Sequential Data," in *Advances in neural information processing systems*, Jun. 2015, pp. 2980–2988.

[17] M. Fraccaro, S. r. K. S nderby, U. Paquet, and O. Winther, "Sequential Neural Models with Stochastic Layers," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016, pp. 2199–2207.

[18] R. G. Brown and P. Y. C. Hwang, "Introduction to Random Signals and Applied Kalman Filtering," p. 3.

[19] A. Doucet and A. M. Johansen, "A Tutorial on Particle Filtering and Smoothing: Fifteen years later."

[20] R. G. Krishnan, U. Shalit, and D. Sontag, "Deep Kalman Filters," in *AAAI Conference on Artificial Intelligence*, Feb. 2017.

[21] C. J. Maddison, D. Lawson, G. Tucker, N. Heess, M. Norouzi, A. Mnih, A. Doucet, and Y. W. Teh, "Filtering Variational Objectives," in *Advances in Neural Information Processing Systems*, May 2017, pp. 6576–6586.

[22] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *arXiv:1609.03499 [cs]*, Sep. 2016, arXiv: 1609.03499.

[23] D. Nguyen, R. Vadaine, G. Hajduch, R. Garello, and R. Fablet, "A Multi-task Deep Learning Architecture for Maritime Surveillance using AIS Data Streams," in *2018 IEEE DSAA*, Oct. 2018.

[24] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, May 2013.

[25] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.