INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Privacy-preserving Variational Information Feature Extraction for Domestic Activity Monitoring Versus Speaker Identification

Alexandru Nelus[1], *Janek Ebbers*[2], *Reinhold Haeb-Umbach*[2], *Rainer Martin*[1]

[1]Ruhr-Universität Bochum, Institute of Communication Acoustics, Bochum, Germany
[2]Universität Paderborn, Fachgebiet Nachrichtentechnik, Paderborn, Germany
{alexandru.nelus, rainer.martin}@rub.de, {ebbers, haeb}@nt.uni-paderborn.de

Abstract

In this paper we highlight the privacy risks entailed in deep neural network feature extraction for domestic activity monitoring. We employ the baseline system proposed in the Task 5 of the DCASE 2018 challenge and simulate a feature interception attack by an eavesdropper who wants to perform speaker identification. We then propose to reduce the aforementioned privacy risks by introducing a variational information feature extraction scheme that allows for good activity monitoring performance while at the same time minimizing the information of the feature representation, thus restricting speaker identification attempts. We analyze the resulting model's composite loss function and the budget scaling factor used to control the balance between the performance of the trusted and attacker tasks. It is empirically demonstrated that the proposed method reduces speaker identification privacy risks without significantly deprecating the performance of domestic activity monitoring tasks.

Index Terms: privacy, variational information, adversarial attack, feature interception, latent variable, mutual information, deep neural networks

1. Introduction

The monitoring of domestic activities in the context of smart home environments is gaining increasing interest, being fueled by the surge in popularity of digital home assistants (e.g., Amazon Echo, Google Home, etc.) and the growing need for Ambient Assisted Living (AAL) solutions [1]. The Detection and Classification of Acoustic Scenes and Events (DCASE) challenge has promoted this field in Task 5 of its 2018 edition [2] by providing an extensive database of domestic sounds derived from the SINS dataset [3] and by asking participants to develop algorithms for their classification. The dataset is based on annotated multi-channel and multi-node recordings from a person living in a vacation home for one week, thus offering a highly informative and spatial-cues-enabled perspective.

The solutions proposed by the challenge's participants have all employed machine learning (ML) approaches, more specifically various deep neural network (DNN) architectures [4]. When considering the deployment of such solutions in real-world distributed acoustic sensor network (ASN) scenarios, e.g., where for network efficiency reasons the DNN-based feature extraction is implemented at sensor level and the classification decision is centralized [5], we also have to take into account the inherent privacy implications. In this regard, we envision a case where the DNN-based feature representation is intercepted during inference by a third-party attacker who wants to use it for a more privacy-invasive task such as speaker identification. Therefore, we aim to determine the resulting privacy risks and we propose to tackle them by employing privacy-preserving variational information feature extraction. This produces a lossy information minimization and the effects thereof are then analyzed w.r.t. domestic activity monitoring and speaker identification performance. Due to the intrinsic relation between some of the speech-based domestic activity classes (e.g., social activity) and the attacker's task, this challenging scenario will not have a trivial solution.

The remainder of this paper is organized as follows: We first discuss the relation to prior work, we then describe the privacy-preserving feature extraction model, followed by a description of the neural network architecture used, after which we detail the experimental layout and the results, finalizing with conclusions and ideas for future work.

2. Relation to prior work

The topic of privacy-preserving feature extraction was initially investigated by the authors in [6], where generative adversarial feature extraction was used in a conflicting goals scenario in order to control the trade-off between gender recognition and speaker identification. Although efficient, this method did not lead to a generalized information minimization technique due to its dependency on a specific attacker configuration. In the same context of gender recognition vs. speaker identification the authors have then proposed a solution inspired by variational information autoencoders [7] where the encoding variable is a compact data representation and where a reparametrization trick [8] is used to allow stochastic sampling during backpropagation. Mutual information (MI) is further chosen as an information regularization criterion for the loss function as supported by works like [9] and [10], where it is successfully used to increase network performance and robustness against adversarial attacks in the testing domain. This solution has produced positive results as indicated in [11] and we consider it to be a suitable starting point for the more complex, domestic activity monitoring vs. speaker identification scenario.

As far as the authors are aware, at the time of writing this paper, there is no previous investigation on using variational information networks against adversarial attacks based on speaker identification in the context of domestic activity monitoring.

3. Trust versus threat model

3.1. System description

The DCASE 2018 challenge offers a baseline system for Task 5 composed of two convolutional layers and one dense layer intended to make task participation easier and to provide reference performance. By employing the taxonomy introduced in [12] we will refer to this as the *trust model* and we will further adapt it to our proposed variational information method and

http://dx.doi.org/10.21437/Interspeech.2019-1703

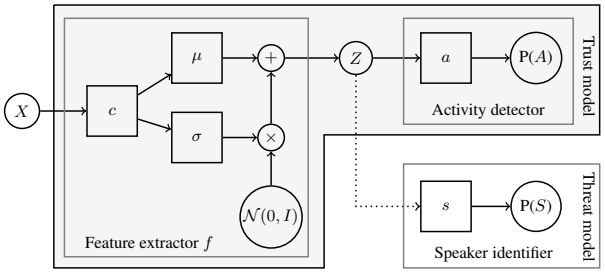

Figure 1: *Flow chart of privacy-preserving feature extraction for domestic activity monitoring vs. speaker identification.*

distributed ASN scenario. We will then analyze the changes in performance in comparison with the baseline performance and with the performance of the speaker identification system. Again employing the taxonomy introduced in [12], we will refer to the latter as the *threat model*.

The trust and threat models are depicted in Fig.1 and are described as follows. The trust model consists of a feature extraction block f which transforms the low-level feature set X into the privacy-preserving feature set Z. The latter is then forwarded to the multilayer perceptron (MLP) based activity detector a with the weights and biases parameters Φ_a which in turn estimates the domestic activity labels' probabilities $P(A)$.

The feature extractor block f comprises of a convolutional neural network (CNN) based structure c with weights and biases parameters Φ_c which uses as input the low-level feature set X. The CNN's output is concomitantly passed to the dense layers μ and σ which have the respective weights and biases parameters Φ_μ and Φ_σ. The output of the σ layer is multiplied with samples from a ζ-dimensional standard normal distribution $\mathcal{N}(0, I)$ and the result is then added to the output of the μ layer, creating thus the privacy-preserving feature set Z. The motivation behind this stochastic encoding will be detailed in section 3.2.

The threat model, consisting of an MLP-based classifier s with weights and biases parameters Φ_s, intercepts the privacy-preserving feature set Z and uses it in order to estimate the speaker labels' probabilities $P(S)$.

3.2. Training the trust model

In the proposed scenario the objective of the trust model is to develop a feature representation that can further be used for domestic activity monitoring but which, at the same time, when intercepted by an attacker can offer little task-extraneous information.

The objective's first part can be separately expressed as minimizing the cross-entropy between the activity labels' true $P(A^t)$ and estimated $P(A)$ probability distributions,

$$\min_{\Phi_c, \Phi_\mu, \Phi_\sigma, \Phi_a} \mathbb{E}_{A^t \sim p(A^t)}[-\log p(A)]. \tag{1}$$

The objective's second part can be separately handled by minimizing the information in the privacy-preserving feature representation Z and thus diminishing its discriminative characteristics. As an information regularization criterion we minimize the mutual information between the low-level and privacy-preserving feature sets $I(X; Z)$, as also supported by works like [9, 10]. Even though this has the advantage of being a comprehensive data similarity measure, it does come with the disadvantage of being computationally demanding. A more computationally practical solution is to find an MI upper bound $I_{max}(X; Z) \geq I(X; Z)$ and use this in the network training process. In this regard we first formulate the entropy-based expression of MI:

$$\begin{aligned} I(X; Z) &= H(Z) - H(Z|X) \\ &= -\int p(z) \log p(z) dz + \int p(x, z) \log p(z|x) dx dz. \end{aligned} \tag{2}$$

An analytical expression of $H(Z|X)$ can be obtained by employing a stochastic encoding mechanism as introduced by [8]. For this we construct a normal-distributed encoding variable $z = \sigma(c(x)) \cdot \epsilon + \mu(c(x))$, where $\epsilon \sim \mathcal{N}(0, I)$. This forces the conditional distribution of z given the input variable x to follow a Gaussian distribution:

$$p(z|x) = \mathcal{N}(\mu(c(x)), \sigma(c(x))). \tag{3}$$

Given that stochastic sampling from $p(z|x)$ during backpropagation is intractable we invoke a reparametrization trick [8] and substitute it with sampling from z. The latter can be regarded as a deterministic variable and its parameters Φ_μ and Φ_σ from layers μ and σ can be updated during backpropagation.

In order to find an analytical expression for $H(Z)$ we introduce a variational distribution $q(z)$ and, as also suggested by [9], we assume it to be standard Gaussian $\mathcal{N}(0, I)$. We then use the Kullback-Leibler divergence's property of always being positive [13] in order to obtain an upper bound for $H(Z)$:

$$-\int p(z) \log p(z) dz \leq -\int p(z) \log q(z) dz. \tag{4}$$

Using (2) and (4) we get that the upperbound of $I(X; Z)$ is

$$\int p(x, z) \log \frac{p(z|x)}{q(z)} dx dz = KL(p(z|x)||q(z)), \tag{5}$$

where KL is the Kullback-Leibler distance between the conditional $p(z|x)$ and variational $q(z)$ distributions. We next define the MI upper bound $I_{max}(X; Z) \geq I(X; Z)$ as:

$$I_{max}(X; Z) = KL(p(z|x)||q(z)). \tag{6}$$

By minimizing $I_{max}(X; Z)$ we can now reduce the dependency between the low-level feature representation X and the privacy-preserving feature representation Z.

Since both $p(z|x)$ and $q(z)$ are Gaussian distributions, the analytical expression of their Kullback-Leibler distance is further employed to obtain

$$I_{max}(X; Z) = \frac{1}{2}\Big(\mathrm{tr}(\Sigma_z) + \mu_z^\top \mu_z - \log\det(\Sigma_z) - \zeta\Big), \tag{7}$$

where $\Sigma_z = \mathrm{diag}(\sigma(c(x))^2)$ and $\mu_z = \mu(c(x))$.

Considering the earlier mentioned trust model's objectives of providing good activity monitoring performance, expressed by (1) while at the same time reducing task-extraneous information by minimizing $I_{max}(X; Z)$, the trust model's loss function to be minimized can be expressed as:

$$\min_{\Phi_c, \Phi_\mu, \Phi_\sigma, \Phi_a} \mathbb{E}_{A^t \sim p(A^t)}[-\log p(A)] + \beta I_{max}(X; Z). \tag{8}$$

The intuitive explanation behind this formulation is that feature extraction will be forced to initially discard MI between X and Z that is irrelevant for domestic activity detection but not at a high cost for the task's performance. Similarly to [6] and [11], a *budget scaling* factor β is introduced to control the balance between how much domestic activity monitoring performance we wish to renounce in favor of a less informative privacy-preserving feature representation.

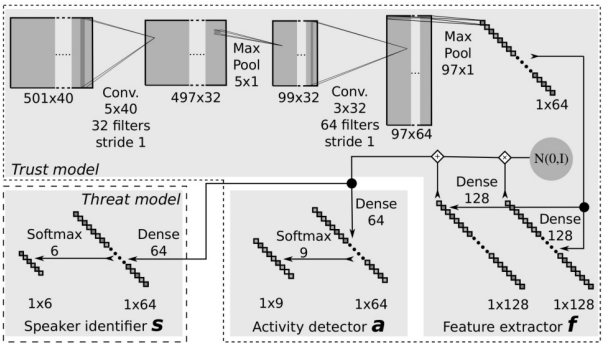

Figure 2: *Network architecture for privacy-preserving variational information feature extraction.*

3.3. Training the threat model

The feature extractor f is employed to extract the privacy-preserving feature representation Z thus simulating an interception attack by the attacker s during trust model inference. The goal of the attacker is to perform speaker identification using the intercepted feature set Z. This is done by minimizing the cross-entropy between the speaker labels' true $P(S^t)$ and estimated $P(S)$ probability distributions and only updating the Φ_s parameters:

$$\min_{\Phi_s} \mathbb{E}_{S^t \sim p(S^t)}[-\log p(S)]. \quad (9)$$

4. Network configuration

4.1. Low-level feature extractor

As indicated in the DCASE 2018 Task 5 baseline system [2], the log mel-band energy (LMBE) representation of the signal $x_s(t)$ is used as network input and is obtained as follows: The short-time Fourier transform (STFT) $X_{\text{stft}}(\kappa, b)$ with window length L_1 and step R_1 is applied to $x_s(t)$, where κ and b denote the frequency bin and time frame index, respectively. The obtained squared-magnitude spectrum is then mapped onto the Mel scale [14], resulting in the Mel-spectrum $X_{\text{mel}}(k', b)$, where $k' = 0, 1, \ldots, K' - 1$ is the index of the Mel scale frequency bin. Finally, by taking the logarithm of the absolute Mel-spectrum we obtain the LMBE features,

$$X_{\text{lmbe}}(k', b) = \log |X_{\text{mel}}(k', b)|. \quad (10)$$

4.2. Privacy-preserving feature extractor

The architecture of the feature extractor f is depicted in Fig. 2 for $K' = 40$, $L_1 = 0.04s$, $R_1 = 0.02s$ and a $x_s(t)$ length of 10s. We use here the CNN-based c block from the DCASE 2018 Task 5 baseline system [2] which consists of two convolutional layers. The first one is of size $501 \times K'$ containing 32 filters with a kernel size of $K' \times 5$ and stride 1, thus convolution is only performed over the time axis. Subsampling is then performed by Max Pooling with a factor of 5. The second CNN layer is of size 99×32 with 64 filters that have a kernel size of 3×32 and a stride of 1. Subsequently, this is subsampled using Global Max Pooling by a factor 97. After each CNN layer Batch Normalization, ReLU activation and Dropout (20%) [15] are used. We now introduce our proposed variational information approach by passing the resulting output of size 1×64 to the dense layers μ and σ, each containing 128 neurons. The output of layer σ is multiplied with samples from a ζ-dimensional standard normal distribution $\mathcal{N}(0, I)$ with $\zeta = 128$ and added to the output of layer μ. The resulting privacy-preserving feature representation has the form 1×128.

Table 1: *Division of DCASE 2018 and WSJ audio data into training and testing sets.*

Set	Activity	Segments	Sessions
Development: 65% train, 35% test	Absence	18860	42
	Cooking	5124	13
	Dishwashing	1424	10
	Eating	2308	13
	Other	2060	118
	Social Activity	4944	21
	Vacuum cleaning	972	9
	Watching TV	18648	9
	Working	18644	33
Set	Distribution	Segments	Balance
WSJ: 80% train 20% test	18 groups 6 spk./group	Avg. 87/spk.	Gender balanced

4.3. Activity detector and speaker identifier

We perform activity detection and speaker identification for each resulting privacy-preserving feature vector Z by using the MLP architectures a and respectively s presented in Fig. 2. Both MLP architectures consist of 64 fully connected nodes that use ReLU activation functions, Dropout (20%) and a final layer of 9 respectively 6 output nodes, on which we apply a Softmax function. The number of output nodes corresponds to the total number of domestic activity classes and respectively to the total number of speakers from our speaker identification pool. In all training phases we employ the Adam optimizer [16] with a learning rate of 0.0001. The activity detector's configuration is identical to the one proposed in the DCASE 2018 Task 5 baseline system.

5. Experiments

5.1. Databases and settings

For domestic activity monitoring we use the DCASE 2018 Task 5 development set [2] which is a subset of the SINS dataset [3]. The data split into training (65%) and testing (35%) files is provided by the challenge's baseline system [17] as well as the arrangement into four folds. For ease of implementation and considering that all folds contain the same files, differing only in their train/test assignation, we have arbitrarily concentrated in this work on fold number 3. The activities considered in this experiment along with the total number of 10 s segments and sessions for each activity class are shown in Table 1.

For simulating a speaker identification attack in a smart home environment we select 108 speakers from the Wall Street Journal (WSJ) corpus [18], of which 54 are male and 54 are female, and we arbitrarily group them into 18 groups where each group is composed of 3 male and 3 female speakers. We then arrange the data into 10 s segments in order to match the network's input shape requirements, thus obtaining an average of 87 segments/speaker. For each speaker we split the segments into training (80%) and testing (20%). The aforementioned specifications are also summarized in Table 1.

The accuracy metric for both domestic activity monitoring and speaker identification performance is, as also used in the DCASE 2018 challenge [2], the macro-averaged F_1-score,

which is the mean of the class-wise F_1-scores:

$$F_1 = \frac{1}{M}\sum_{m=1}^{M}\frac{2P(m)R(m)}{P(m)+R(m)}, \text{where}$$

$$P(m) = \frac{TP(m)}{TP(m)+FP(m)} \qquad R(m) = \frac{TP(m)}{TP(m)+FN(m)}$$

with m being the domestic activity/speaker class, M the total number of activity/speaker classes and $TP(m)$ the number of true-positives, $FP(m)$ the number of false-positives, and $FN(m)$ the number of false-negatives for class m.

5.2. Domestic activity monitoring

In order to obtain a reference level for domestic activity monitoring performance we first employ the original baseline system downloaded from [17]. The system is implemented in Python, and uses the DCASE UTIL library [19] for dataset handling and low-level feature extraction and the Keras library [20] for neural network operations. We train the network on fold number 3 with the settings specified in 5.1 using a batch size of 256 input segments for 400 epochs. Each 10 epochs the model is saved and in the end the best performing one is chosen for testing. The obtained macro-averaged reference F_1-score is of 83.80% and is shown in Fig. 3 under the indicator "Reference". The individual F_1-scores for each class are also indicated.

We then adapt the baseline system by adding the variational information components as indicated in section 3 and we systematically vary the budget scaling factor β throughout this experiment. For each value of β we train the network on fold number 3 with the settings specified in 5.1, again using a batch size of 256 input segments for 400 epochs. As previously, each 10 epochs the model is saved and in the end the best performing one is chosen for testing. The obtained macro-averaged F_1-scores along with the individual F_1-scores for each class are depicted in Fig. 3 under their corresponding β values.

5.3. Speaker identification attack

We next train the attacker part of our proposed system by concatenating the already trained feature extractor f with the attacker architecture detailed in Fig. 2 for each value of β. We train on the WSJ dataset with the settings specified in section 5.1 using a batch size of 256 input segments for a maximum of 1000 epochs or until interruption by a cross-entropy-based early stopping function [21] that prevents overfitting. This experiment is performed for 10 cross-validation sessions and the macro-averaged F_1-scores for each 6-speakers group are averaged across all 18 groups and the results are again averaged over the 10 cross-validation sessions. The obtained results are depicted in Fig. 3 under the label "Attacker" for each corresponding β value. The figure also shows a control line for random guessing speaker identification F_1-score.

5.4. Discussion

Our first observation is that the feature set produced by the baseline system, when intercepted by an attacker, allows for a speaker identification F_1-score of 47.07%. When considering the deployment to a smart-home environment this poses substantial privacy risks. Once we replace the deterministically-produced baseline feature set with a stochastically-produced feature representation the attacker's performance drops to 29.40% while domestic activity performance has just an insignificant drop in performance, resulting in an F_1-score of

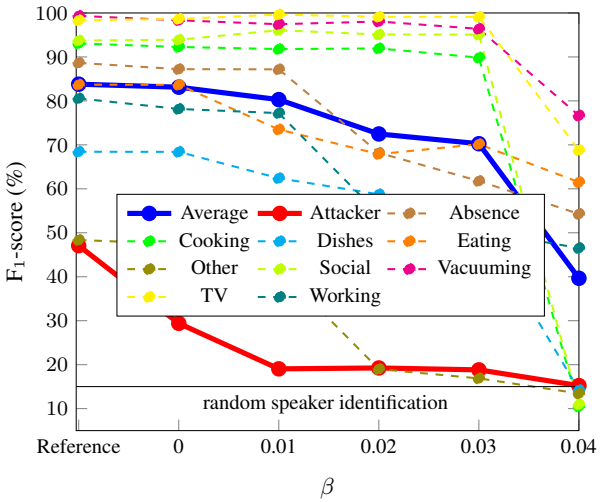

Figure 3: *The influence of the budget scaling factor β on domestic activity monitoring and speaker identification attacks.*

83.12%. Note that at this point $\beta = 0$ thus no minimization of $I_{max}(X;Z)$ is yet performed and the drop in attacker performance is only due to the stochastic sampling from $N(0,I)$ of the latent variable Z, hinting that a class-wise-generic feature representation greatly reduces class-extraneous information.

Further on, we observe that by minimizing $I_{max}(X;Z)$ with $\beta = 0.01$ we obtain an even lower attacker F_1-score of 19.09% which is now close to random guessing values. The loss in activity monitoring performance is small, resulting in an F_1-score of 80.29%. When we continue to increase the value of β, thus increasing the weight assigned to minimizing $I_{max}(X;Z)$, we observe that speaker identification performance keeps deprecating until random guessing values are reached but now the impact on activity monitoring performance is significantly higher. What is interesting to observe is that the performance of speech related activities e.g., "Social activity" is rather resilient to increases of β, pointing towards the possibility of further combining the current method with additional countermeasures specifically targeted at speaker identification such as adversarial training [6].

6. Conclusions and future work

We have highlighted the privacy risks entailed by DNN-based feature representations in the context of a distributed ASN for domestic activity monitoring. Empirical evidence was provided to show that variational information feature extraction can be successfully used to drastically reduce the effects of speaker identification attacks on intercepted features without significantly altering activity monitoring performance. For this, a privacy-preserving latent feature representation along with a general loss function were proposed and a budget scaling factor was introduced and analyzed. In future work we aim to supplement the information minimization measure with a specifically-targeted attacker countermeasure such as adversarial training and to also consider non-parametric MI estimation.

7. Acknowledgements

This work has been supported by DFG under contract no. Ma 1769/6-1 and Ha 3455/15-1 within the Research Unit FOR 2457 (Acoustic Sensor Networks).

8. References

[1] F. Erden, S. Velipasalar, A. Alkar, and A. Cetin, "Sensors in assisted living: A survey of signal and image processing methods," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 36–44, 3 2016.

[2] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics," KU Leuven, Tech. Rep., 2018. [Online]. Available: https://arxiv.org/abs/1807.11246

[3] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 32–36.

[4] DCASE 2018. DCASE2018 results: Monitoring of domestic activities based on multi-channel acoustics. [Online]. Available: http://dcase.community/challenge2018/task-monitoring-domestic-activities-results

[5] H. Afifi, S. Auroux, and H. Karl, "Marvelo: Wireless virtual network embedding for overlay graphs with loops," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[6] A. Nelus and R. Martin, "Gender discrimination versus speaker identification through privacy-aware adversarial feature extraction," in *Speech Communication; 13. ITG Symposium; Proceedings of*. VDE, 2018.

[7] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[9] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[10] Y. Zhang, M. Ozay, Z. Sun, and T. Okatani, "Information potential auto-encoders," *arXiv preprint arXiv:1706.04635*, 2017.

[11] A. Nelus and R. Martin, "Privacy-aware feature extraction for gender discrimination versus speaker identification," in *International Conference on Acoustics, Speech, and Signal Processing; Proceedings of*. IEEE, 2019.

[12] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "Sok: Security and privacy in machine learning," in *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2018, pp. 399–414.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience, 2006.

[14] S. Furui, *Digital speech processing: synthesis, and recognition*. CRC Press, 2000.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] DCASE. (2018) DCASE2018: Monitoring of domestic activities based on multi-channel acoustics. [Online]. Available: http://dcase.community/challenge2018/task-monitoring-domestic-activities/

[18] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the Workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[19] T. Heittola. (2018) DCASE UTIL: Utilities for detection and classification of acoustic scenes. [Online]. Available: https://dcase-repo.github.io/dcase util/

[20] Keras. (2018) Keras: The python deep learning library. [Online]. Available: https://keras.io/

[21] L. Prechelt, "Automatic early stopping using cross validation: quantifying the criteria," *Neural Networks*, vol. 11, no. 4, pp. 761 – 767, 1998.