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

Privacy-preserving Siamese Feature Extraction for Gender Recognition Versus Speaker Identification

Alexandru Nelus, Silas Rech, Timm Koppelmann, Henrik Biermann, and Rainer Martin

Ruhr-Universität Bochum, Institute of Communication Acoustics, Bochum, Germany
{alexandru.nelus, silas.rech, timm.koppelmann, henrik.biermann, rainer.martin}@rub.de

Abstract

In this paper we propose a deep neural-network-based feature extraction scheme with the purpose of reducing the privacy risks encountered in speaker classification tasks. For this we choose a challenging scenario where we wish to perform gender recognition but at the same time prevent an attacker who has intercepted the features to perform speaker identification. Our approach is to employ Siamese training in order to obtain a feature representation that minimizes the Euclidean distance between same gender speakers while maximizing it for different gender speakers. It is experimentally shown that the obtained effect is of anonymizing speakers from the same gender class and thus drastically reducing privacy risks while still permitting class discrimination with a higher accuracy than other previously investigated methods.

Index Terms: privacy, utility, machine learning, privacy preservation, privacy attack, inference attack, trust model, threat model, feature extraction

1. Introduction

Recent advances in machine learning (ML) coupled with the expansion of computational resources have propelled ML-based solutions to an omnipresent component of modern day devices and services. Besides the obvious contribution to fostering multidisciplinary progress, ML and associated data collection efforts also entail inherent privacy and security risks, the study of which, albeit nascent, has gathered increasing importance [1].

This paper will exemplify the privacy risks resulting from the use of advanced ML methods for audio signal classification by proposing a challenging example based on conflicting goals. We address the issue of an attacker intercepting features extracted using a deep neural network (DNN) for the task of speaker gender recognition and trying to use them for a more privacy-invasive task like speaker identification. Recognizing the gender of a speaker has received increasing attention in the context of gender equality measures [2, 3] where DNN-based solutions are considered state of the art [4]. At the same time the General Data Protection Regulation [5] encourages the use of *privacy by design* when handling personal data [6], thus motivating our current approach.

As mentioned in [7], a systems's privacy can be measured w.r.t. the adversarial goals that it was designed to defend against, thus we will regard privacy here as being related to the performance of speaker identification experiments. By employing the privacy taxonomy terms introduced by [7], we can frame our proposed scenario as follows: the *trust model* is a DNN designed to perform speaker gender recognition; the *threat model* or *attacker* is a DNN designed to perform speaker identification using intercepted features; as the envisioned attack happens after the network has been already trained and deployed, we will refer to it as an *inference attack* through feature interception; we consider this to be a *white-box* attack where the algorithm of the trust model is public. Previous work [8] has indicated that a feature representation suitable for gender recognition also carries a significant amount of speaker dependent information. Considering this, we propose to use Siamese training [9] to obtain a privacy-preserving feature representation that serves the aforementioned gender recognition task and at the same time protects against speaker identification attacks.

The remainder of this paper is organized as follows: We discuss the relation to prior work, we afterwards describe the trust and threat models including the Siamese training procedure, followed by a description of the neural network architecture used, the experimental layout and wrapping up with conclusions.

2. Relation to prior work

The concept of privacy-preserving feature extraction has been investigated by the authors in [8] where adversarial training was used to control the trade-off between gender recognition and speaker identification. An attacker-independent, more generalized approach was further pursued in [10] where variational information was used in order to limit the amount of information exposed by the feature representation. Although both approaches provided good results, they failed to completely remove all speaker identity characteristics without strongly impacting gender recognition. One possible factor may be the insufficient degrees of freedom of the feature representation as indicated by [11].

Our proposed solution is inspired by Siamese networks which were first introduced in the framework of signature verification [9] and which function by forcing the feature representation of similar input variables to lie in close Euclidean proximity while distancing dissimilar input variable representations. This concept has been recently used by [12] in the context of privacy-preserving video processing for mobile analytics. As far as the authors are aware, at the time of writing this paper, there is no previous investigation on using Siamese training for privacy-preserving audio feature extraction.

3. Trust versus threat model

3.1. System description

We first present an overview of the proposed system after which, in the following subsections, we characterize its constituting parts. As also depicted in Figure 1, the trust model consists of a feature extraction block f and a gender recognizer block g. The former transforms the low-level feature representation X, expressed by feature vectors x into the high-level feature representation Z expressed by feature vectors z. The feature extraction block f is composed of a number of G convolutional neural network (CNN) groups c_j with weights and biases parameters Φ_c, where $j \in [1, G]$. Each CNN group c_j is in turn composed

http://dx.doi.org/10.21437/Interspeech.2019-1148

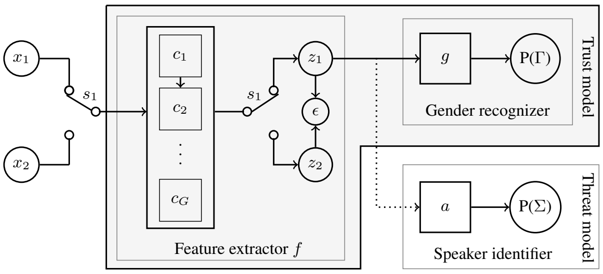

Figure 1: *Flow chart of Siamese privacy-preserving feature extraction for gender recognition vs. speaker identification.*

of a batch normalization layer followed by three convolutional layers and ending with a max pooling layer as further detailed in section 4. During Siamese feature extraction, pairs of low-level feature vectors (x_1, x_2) are used as input in an intra-pair sequential fashion. This is indicated in Figure 1 by the s_1 switch which first connects x_1 to z_1 and then x_2 to z_2. The Euclidean distance ϵ between two resulting high-level feature vectors z_1 and z_2 is then computed and used in the network's loss function.

After the Siamese feature extraction is completed, the already trained feature extractor f is used in conjunction with the multilayer perceptron (MLP) gender recognizer g, with weights and biases parameters Φ_g, in order to estimate the speaker gender label's probabilities $P(\Gamma)$. The weights and biases parameters Φ_c of the feature extractor f are kept fixed and solely the weights and biases parameters Φ_g of g are trained. Only one low-level input vector x is required for gender recognition. This is also depicted in Figure 1 when considering the use of the x_1 to z_1 path exclusively.

The threat model (attacker), which consists of an MLP-based classifier a with weights and biases parameters Φ_a, intercepts the high-level feature vectors z during gender recognition inference, and uses them in order to estimate the speaker label's probabilities $P(\Sigma)$.

3.2. Feature extractor

In our proposed scenario, the objective of the feature extractor is to develop a feature representation that can still be used to discriminate between male and female speakers but which, at the same time, can anonymize the identity of the speakers. Inspired by the concept of *k-anonymity* [13] where the information of each individual contained in a dataset should not be distinguishable from at least k-1 other individuals from the same dataset, we propose separately mapping all female and all male speakers into very close class-wise Euclidean proximity representations. In this way, an attacker will have a more difficult task in identifying one specific female or male speaker out of a group of female or male speakers. At the same time, we still wish to have an easy discrimination between the male and female groups and to achieve this, we propose to increase the Euclidean distance between the feature mappings of the two gender classes. This so far is exactly the principle behind Siamese neural networks, which is our proposed approach for achieving the aforementioned objectives.

The traditional layout of Siamese networks as introduced by [9] consists of two identical neural networks with conjoined (shared) weights and biases which concomitantly process two input vectors and optimize their Euclidean distance. For ease of implementation we have translated this parallel structure into a single neural network, the feature extractor f, where the two input vectors are now processed sequentially and their Euclidean distance is then optimized. The effects are obviously identical. We refer to the two low-level feature vector inputs as being *similar* if they belong to the same gender class (e.g., from two male or two female speakers) and *dissimilar* if they belong to different gender classes (e.g., from one male and one female speaker). The contrastive loss function to be minimized upon training is:

$$\min_{\Phi_c} \mathbb{E}_{Y\sim p(Y)} \left[(1-Y)\epsilon^2 + Y(\max(0, marg-\epsilon)^2)\right] \quad (1)$$

where $\epsilon = \|z_1 - z_2\|_2$ is the Euclidean distance between two resulting high-level feature vectors z_1 and z_2, and Y is a binary variable indicating whether the input pair (x_1, x_2) is similar or dissimilar. The maximum inter-class Euclidean distance is controlled by $marg$. In traditional Siamese network implementations, during inference, ϵ is compared with a $marg/2$ threshold:

$$\begin{cases} \epsilon \le marg/2 & \text{then } z_1 \text{ and } z_2 \text{ are similar} \\ \epsilon > marg/2 & \text{then } z_1 \text{ and } z_2 \text{ are dissimilar} \end{cases} . \quad (2)$$

As we are interested in identifying the feature representation's non-linear discriminative characteristics and we also wish to obtain a class prediction for a single input vector thus avoiding the need for input pairs, we will only use the above thresholding for validation purposes as further described in section 5. The gender classification task, using the now trained Siamese feature extractor f, will be performed by the gender recognizer network described next.

3.3. Gender recognizer

The objective of the gender recognizer g is to perform speaker gender recognition using the feature representation provided by the feature extractor f. This is implemented in a separate training procedure where we use the entire trust model depicted in Figure 1 while keeping Φ_c fixed and only updating Φ_g. This can be formulated as as minimizing the cross-entropy between the gender labels' true $P(\Gamma^t)$ and estimated $P(\Gamma)$ probability distributions:

$$\min_{\Phi_g} \mathbb{E}_{\Gamma^t\sim p(\Gamma^t)}[-\log p(\Gamma)]. \quad (3)$$

3.4. Speaker identifier

We use the feature extractor f to extract the high-level feature representation Z which is then used by the attacker (threat model). The goal of the attacker is to perform speaker identification as best as possible using the intercepted feature set Z. This is implemented in a separate training procedure where we concatenate f with a and only update Φ_a while keeping Φ_c fixed. The objective loss function to be minimized is the cross-entropy between the speaker labels' true $P(\Sigma^t)$ and estimated $P(\Sigma)$ probability distributions:

$$\min_{\Phi_a} \mathbb{E}_{\Sigma^t\sim p(\Sigma^t)}[-\log p(\Sigma)]. \quad (4)$$

4. Network configuration

4.1. Low-level feature extraction

We use the mel-frequency cepstral coefficient (MFCC) representation of the signal $x_{sig}(t)$ as the low-level feature input for the neural-network-based high-level feature extractor f. After applying a short-time Fourier transform (STFT) $X_{\text{stft}}(\kappa, b)$ with window length L_1 and step R_1 to $x_{sig}(t)$, where κ and b denote the frequency bin and time frame index, respectively,

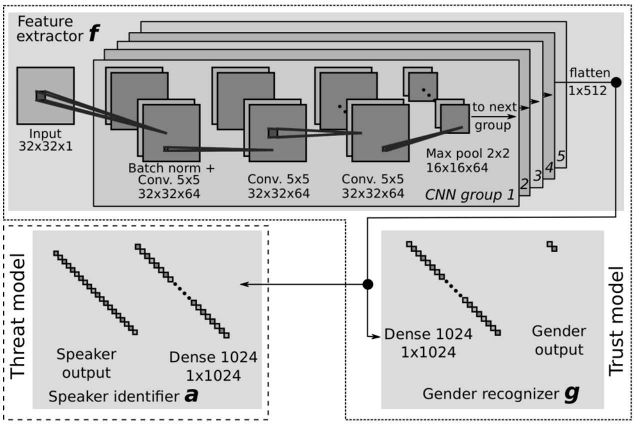

Figure 2: *Network architecture for Siamese privacy-preserving feature extraction.*

we map the squared-magnitude spectrum 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. The MFCC features $X_{\text{mfcc}}(\eta, b)$ with the cepstral coefficient index $\eta = 0, 1, \ldots, K'' - 1$ are computed by taking the discrete cosine transform of the logarithm of the absolute Mel-spectrum and selecting the first K'' coefficients.

4.2. High-level Siamese feature extraction

The architecture of the Siamese feature extractor f is inspired by the VGG-16 layout [15] and shown in Figure 2 for $G = 5$ CNN-based c_j groups, where $j \in [1, G]$. Each group c_j consists of a batch normalization layer followed by a series of three identical convolutional layers of sizes $\left[\frac{K''}{2^{j-1}} \times \frac{K''}{2^{j-1}}\right]$containing F_j kernels of size F'_j, and ends with a max pooling layer of stride 2 and filter size $[2 \times 2]$. The feature extractor uses the low-level feature stream X_{mfcc} of the form $\left[\frac{T}{R_1 K''} \times K'' \times K''\right]$, where T is the signal's time length. The extractor's output is stacked in the form of $\left[\frac{T}{R_1 K''} \times \frac{K''}{2^G} \times \frac{K''}{2^G} \times F_G\right]$ and then flattened, where each resulting high-level feature vector is responsible for a receptive field of length $K'' R_1$ s. In Figure 2 this is depicted for the empirically chosen values of $K' = 64$, $K'' = 32$, $(F_1, F_2, F_3, F_4, F_5) = (64, 128, 256, 512, 512)$, $(F'_1, F'_2, F'_3, F'_4, F'_5) = (5, 3, 3, 2, 2)$, $L_1 = 0.026$ s and $R_1 = 0.013$ s. These are further used in section 5.

4.3. Gender recognizer and speaker identifier

We perform gender recognition and speaker identification for each resulting high level feature vector z by using the MLP architectures g and respectively a presented in Figure 2. Both MLP architectures consist of 1024 fully connected nodes that use ReLu activation functions and a final layer of two respectively S_t output nodes, on which we apply a softmax function. For training we employ the Adam optimizer [16] with a learning rate of 0.0002 and we also use a dropout rate of 0.5 [17].

Table 1: *Division of audio data into training, evaluation and testing sets, along with corresponding subsets.*

420 speakers			
20 speakers		400 speakers	
70 % data/ speaker	30 % data/ speaker	Gender-testing set	
Gender-training set	Gender-evaluation set	Attacker-testing set: 20 batches of 20 speakers	
		70 % data/speaker	30 % data/speaker
		Attacker-training subset	Attacker-evaluation subset

5. Experiments

5.1. Database and settings

The database contains 420 speakers from the TIMIT corpus [18], of which 290 are male and 130 are female, with an average of 31 seconds of audio per speaker. From these we build training, evaluation and testing sets. We select $S_t/2$ male and $S_t/2$ female speakers, and randomly split every individual's audio data into *gender-training* (70%) and *gender-evaluation* (30%) sets. We use all the audio data from the remainder $420 - S_t$ speakers as the *gender-testing* set.

The remainder $420 - S_t$ speakers are also divided into subgroups of S_t, and every speaker's audio data is randomly split into *attacker-training* (70%) and *attacker-evaluation* (30%) subsets, together forming the *attacker-testing* set. This structure is also described in Table 1, for $S_t = 20$ speakers as chosen for this implementation. Other parameters used follow the initialization described in section 4.

We propose to use *accuracy* as a performance measure for both gender and speaker classification, where:

$$\textit{accuracy} = \frac{\textit{no. of correctly classified samples}}{\textit{total no. of samples}}. \quad (5)$$

The experiments described in this section are performed for 10 cross-validation iterations, each iteration employing random data shuffling, and the presented accuracies are the averaged values of the aforementioned cross-validation iterations.

5.2. Siamese feature extraction

We first train the feature extractor f using the parameter configuration indicated in section 4.2. During this experiment we systematically vary the total number G of CNN-based groups, where $G \in [1, 5]$, as we have further observed that a larger interval offers no significant change in performance.

For the Siamese training procedure, we group all the low-level features X into pairs (x_1, x_2) so that in the gender-training set each feature vector is paired with a similar or a dissimilar feature vector. In the end, the ratio of similar and dissimilar pairs is roughly 50/50. In this experiment, the margin parameter $marg$ was empirically set to 1. The Siamese training is performed using a batch size of 150 samples for a maximum of 75 epochs or until it is interrupted by a cross-entropy-based *early stopping* function that prevents overfitting. This uses data from the gender-training set along with the thresholding function described in Eq. 2 to predict the similarity between two input vectors.

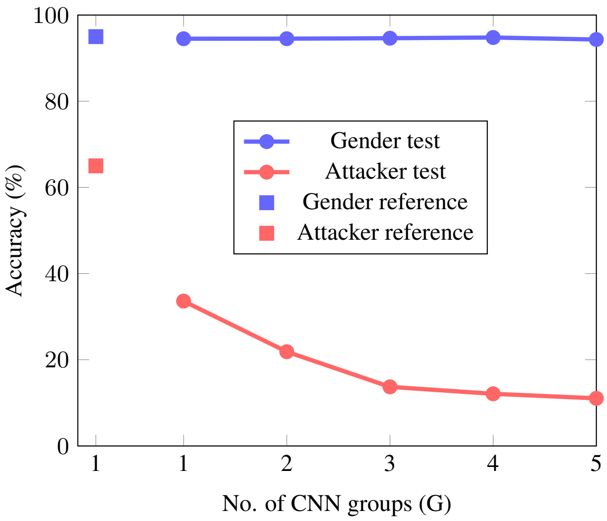

Figure 3: *The influence of the number of CNN groups G on gender recognition (Gender test) vs. speaker identification accuracies (Attacker test) using Siamese feature extraction along with reference accuracies for non-Siamese feature extraction.*

5.3. Gender recognition vs. speaker identification

For each value of G we deploy and analyze the performance of the trust and threat models and show them in Figure 3. The horizontal axis indicates the value of G. We use the already trained feature extractor f and concatenate it with the gender recognizer g as indicated in section 3.3. The gender training is performed for 30 epochs using individual samples (not pairs) from the gender-training grouped in batches of 200 samples. The model is evaluated using individual samples from the gender-evaluation set. In Figure 3, under the label ”Gender test”, we show the gender recognition accuracies using the testing set.

We simulate a feature interception attack by using the already trained feature extractor f and concatenating it with the speaker identifier a as indicated in section 3.4. The network is trained using samples from the attacker-testing set, more specifically from the attacker-training subset. Training is performed for 150 epochs using a batch size of 90 samples. The accuracy of the attack is evaluated using samples from the attacker-evaluation subset and shown in Figure 3, under the label ”Attacker test”.

In order to underline the ML privacy risks and set a point of reference, we have also simulated a non-privacy-preserving gender recognition that uses the trust model with $G = 1$ and which is trained as a traditional CNN network (no Siamese training) using individual low-level input vectors (not pairs) from the gender-training set. During the training procedure all of the trust model's parameters Φ_c and Φ_g are updated. As before, the attacker intercepts the features and uses them for speaker identification. The afferent results are show in Figure 3 under the ”Gender reference” and ”Attacker reference” labels.

5.4. Discussion

In the previously mentioned reference experiment, where no Siamese training is used, the gender recognition accuracy is of 95 % while the speaker identification accuracy of an attacker who has intercepted the features is as high as 65 %, thus illustrating the privacy risks entailed by non-privacy-preserving ML algorithms. After employing privacy-preserving Siamese feature extraction, the attacker accuracy drops to 33 % with minimum impact on the trust model's performance. Further on, we increase the degree of specialization of the feature extractor by increasing the number of CNN groups G and their respective number of kernels F_j while at the same time decreasing their kernel sizes F'_j as indicated in section 4.2. This allows the extractor to gradually progress from larger-scoped to more task-specific features that better serve the Euclidean space optimization. Top performance is achieved for $G = 5$ groups, with a gender testing accuracy of 94 % and an attacker accuracy of 11 %, the latter being very close to the 10 % random guessing attacker accuracy for $S_t = 20$ speakers where the speaker gender is known. We have compared these results with results from our previous investigations that used adversarial [8] and variational-information [10] feature extraction in similar experimental conditions, testing 20 speaker groups from the TIMIT database. This comparison is depicted in Figure 4, where the largest trade-off between gender recognition and speaker identification is selected from the previous works. We argue that the current method's advantage over the former ones lies in specifically including the contrastive Euclidean distance constraints in the training procedure thus anonymizing each gender class separately, while both former approaches tend to anonymize the entire feature set regardless of the gender class.

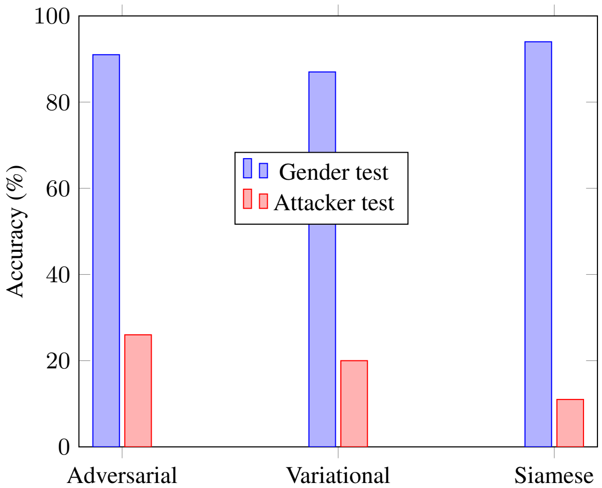

Figure 4: *Comparison of gender recognition (Gender test) vs. speaker identification accuracies (Attacker test) for adversarial [8], variational-information [10] and Siamese feature extraction methods using similar experimental conditions.*

6. Conclusions

Empirical evidence was provided to underline the privacy risks entailed by DNN-based feature extraction and a Siamese network architecture was successfully employed to drastically reduce the aforementioned risks. The degree of network specialization was systematically analyzed and the proposed solution was proven to outperform previously investigated privacy-preserving methods. The system's two competing tasks were chosen as to better illustrate the proposed concept and in future works we intend to expand the investigation to multi-class application scenarios.

7. Acknowledgements

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

8. References

[1] W. House, "Preparing for the future of artificial intelligence," *Executive Office of the President, National Science and Technology Council, Committee on Technology*, 2016.

[2] Gender Equality Commission, "Handbook on the implementation of Recommendation CM/Rec (2013) 1 of the Committee of Ministers of the Council of Europe on Gender Equality and Media," *Council of Europe*, 2015.

[3] Macharia et al., "Who Makes the News?: Global Media Monitoring Project 2015," *World Association for Christian Communication*, 2015.

[4] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5214–5218.

[5] European Parliament and Council, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," 2016.

[6] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding," in *Interspeech 2019*, 2019.

[7] 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.

[8] A. Nelus and R. Martin, "Gender discrimination versus speaker identification through privacy-aware adversarial feature extraction," in *Speech Communication; 13th ITG-Symposium*, Oct 2018, pp. 1–5.

[9] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," in *Advances in neural information processing systems*, 1994, pp. 737–744.

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

[11] Y. Zhang, M. Ozay, Z. Sun, and T. Okatani, "Information potential auto-encoders," *CoRR*, vol. abs/1706.04635, 2017. [Online]. Available: http://arxiv.org/abs/1706.04635

[12] S. A. Ossia, A. S. Shamsabadi, A. Taheri, H. R. Rabiee, N. D. Lane, and H. Haddadi, "A hybrid deep learning architecture for privacy-preserving mobile analytics," *CoRR*, vol. abs/1703.02952, 2017. [Online]. Available: http://arxiv.org/abs/1703.02952

[13] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

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

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[17] 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.

[18] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.