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

Learning Speaker Representations with Mutual Information

Mirco Ravanelli[1], *Yoshua Bengio*[1,2]
[1]Mila, Université de Montréal, Canada
[2]CIFAR, Canada
mirco.ravanelli@gmail.com

Abstract

Learning good representations is of crucial importance in deep learning. Mutual Information (MI) or similar measures of statistical dependence are promising tools for learning these representations in an unsupervised way. Even though the mutual information between two random variables is hard to measure directly in high dimensional spaces, some recent studies have shown that an implicit optimization of MI can be achieved with an encoder-discriminator architecture similar to that of Generative Adversarial Networks (GANs).

In this work, we learn representations that capture speaker identities by maximizing the mutual information between the encoded representations of chunks of speech randomly sampled from the same sentence. The proposed encoder relies on the SincNet architecture and transforms raw speech waveform into a compact feature vector. The discriminator is fed by either positive samples (of the joint distribution of encoded chunks) or negative samples (from the product of the marginals) and is trained to separate them.

We report experiments showing that this approach effectively learns useful speaker representations, leading to promising results on speaker identification and verification tasks. Our experiments consider both unsupervised and semi-supervised settings and compare the performance achieved with different objective functions.

Index Terms: Deep Learning, Speaker Recognition, Mutual Information, Unsupervised Learning, SincNet.

1. Introduction

Deep learning has shown remarkable success in numerous speech tasks, including speech recognition [1–4] and speaker recognition [5, 6]. The deep learning paradigm aims to describe data by means of a hierarchy of representations, that are progressively combined to model higher level abstractions [7]. Most commonly, deep neural networks are trained in a supervised way, while learning meaningful representations in an unsupervised fashion is more challenging but could be useful especially in semi-supervised settings.

Several approaches have been proposed for deep unsupervised learning in the last decade. Notable examples are deep autoencoders [8], Restricted Boltzmann Machines (RBMs) [9], variational autoencoders [10] and, more recently, Generative Adversarial Networks (GANs) [11]. GANs are often used in the context of generative modeling, where they attempt to minimize a measure of discrepancy between a distribution generated by a neural network and the data distribution. Beyond generative modeling, some works have extended this framework to learn features that are invariant to different domains [12] or to noise conditions [13]. Moreover, we recently witnessed some remarkable attempts to learn unsupervised representations by minimizing or maximizing Mutual Information (MI) [14–17]. This measure is a fundamental quantity for estimating the statistical dependence between random variables and is defined as the Kullback-Leibler (KL) divergence between the joint distribution over these random variables and the product of their marginal distributions [18]. As opposed to other metrics, such as correlation, MI can capture complex non-linear relationships between random variables [19]. MI, however, is difficult to compute directly, especially in high dimensional spaces [20]. The aforementioned works found that it is possible to maximize or minimize the MI within a framework that closely resembles that of GANs. Additionally, [15] has further proved that it is even possible to explicitly compute it by exploiting its *Donsker-Varadhan* bound.

Here we attempt to learn good speaker representations by maximizing the mutual information between two encoded random chunks of speech sampled from the same sentence. Our architecture employs both an encoder, that transforms raw speech samples into a compact feature vector, and a discriminator. The latter is alternatively fed by samples from the joint distribution (i.e. two local encoded vectors randomly drawn from the same speech sentence) and from the product of marginal distributions (i.e, two local encoder vectors coming different utterances). The discriminator is jointly trained with the encoder to maximize the separability of the two distributions. We called our approach *Local Info Max (LIM)* to highlight the fact that it is simply based on randomly sampled local speech chunks. Our encoder is based on SincNet [21,22], which efficiently processes the raw input waveforms with learnable band-pass filters based on sinc functions.

The experimental results show that our approach learns useful speaker features, leading to promising results on speaker identification and verification tasks. Our experiments are conducted in both unsupervised and semi-supervised settings and compare different objective functions for the discriminator. We release the code of LIM within the PyTorch-Kaldi toolkit [23].

2. Speaker Representations based on MI

The mutual information between two random variables z_1 and z_2 is defined as follows:

$$\begin{aligned} MI(z_1, z_2) &= \int_{z_1}\int_{z_2} p(z_1, z_2) log\Big(\frac{p(z_1, z_2)}{p(z_1)p(z_2)}\Big) dz_1 dz_2 \\ &= D_{KL}\big(p(z_1, z_2)||p(z_1)p(z_2)\big), \end{aligned} \quad (1)$$

where D_{KL} is the Kullback-Leibler (KL) divergence between the joint distribution $p(z_1, z_2)$ and the product of marginals $p(z_1)p(z_2)$. The MI is minimized when the random variables z_1 and z_2 are statistically independent (i.e., the joint distribution is equal to the product of marginals) and is maximized when the two variables contain the same information (in which case the MI is simply the entropy of any one of the variables).

Our LIM system, depicted in Fig.1, aims to derive a compact representation z. The encoder f_Θ, with $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$, is

http://dx.doi.org/10.21437/Interspeech.2019-2380

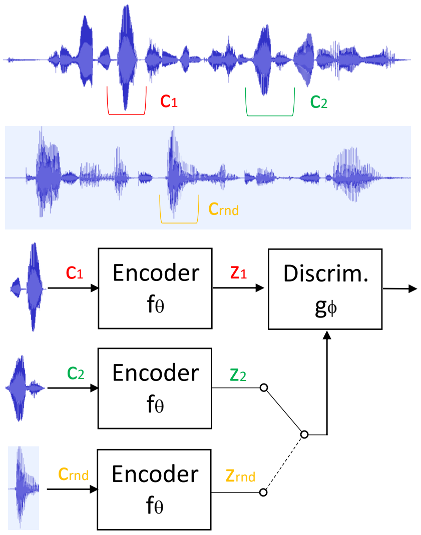

Figure 1: *Architecture of the proposed system for unsupervised learning of speaker representations. The speech chunks c_1 and c_2 are sampled from the same sentence, while c_{rand} is sampled from a different utterance.*

fed by N speech samples and outputs a vector composed of M real values, while the discriminator g_Φ, with $g : \mathbb{R}^{2M} \rightarrow \mathbb{R}$, is fed by two speaker representations and outputs a real scalar. We learn the parameters Θ and Φ of the encoder and the discriminator such that we maximize the mutual information $MI(z_1, z_2)$:

$$(\hat{\Theta}, \hat{\Phi}) = \arg\max_{\Theta,\Phi} MI(z_1, z_2), \tag{2}$$

where the two representations z_1 and z_2 are obtained by encoding the speech chunks c_1 and c_2 that are randomly sampled from the same sentence. Note that one reliable factor that is shared across chunks within each utterance is the speaker identity. The maximization of $MI(z_1, z_2)$ should thus be able to properly disentangle this constant factor from the other variables (e.g., phonemes) that characterize the speech signal but are not shared across chunks of the same utterance.

As shown in Alg. 1, the maximization of MI relies on a sampling strategy that draws positive and negative samples from the joint and the product of marginal distributions, respectively. As discussed so far, the positive samples (z_1, z_2) are simply derived by randomly sampling speech chunks from the same sentence. Negative samples (z_1, z_{rnd}), instead, are obtained by randomly sampling from another utterance. The underlying assumptions considered here are the following: (1) two random utterances likely belong to different speakers, (2) each sentence contains a single speaker only. Under these assumptions, that naturally hold in most of the available speech datasets, our method can be regarded as unsupervised (or, more precisely, self-supervised) because no speaker labels are explicitly used.

A set of N_{samp} positive and negative examples is sampled to form a minibatch $X = \{X_p, X_n\}$. The minibatch X feeds the discriminator g_Φ, that is jointly trained with the encoder. Given z_1, the discriminator g_Φ has to decide whether its other input (z_2 or z_{rnd}) comes from the same sentence or from a different one (and generally a different speaker). Differently to the GAN framework, the encoder and the discriminator are not adversarial here but must cooperate to maximize the discrepancy between the joint and the product of marginal distributions. In other words, we play a *max-max* game rather than a *min-max* one, making it easier to monitor the progress of training (compared to GAN training), simply as the average loss of the discriminator.

Algorithm 1 Learning speaker representation with MI

```
while Not Converged do
    for i=1 to $N_{samp}$ do
        Sample a chunk $c_1$ from a random utterance.
        Sample another chunk $c_2$ from the same utterance.
        Sample a chunk $c_{rnd}$ from another utterance.
        Process the chunks with the encoder:
        $z_1 = f_\Theta(c_1)$, $z_2 = f_\Theta(c_2)$, $z_{rnd} = f_\Theta(c_{rnd})$.
        Create positive and negative samples:
        $X_p[i]$=($z_1$,$z_2$), $X_n[i]$=($z_1$,$z_{rnd}$).
    Compute discriminator outputs: $g(X_p)$, $g(X_n)$.
    Compute Loss $L(\Theta, \Phi)$.
    Compute Gradients $\frac{\partial L}{\partial \Theta}$, $\frac{\partial L}{\partial \Phi}$.
    Update $\Theta$ and $\Phi$ to maximize L.
```

Different objectives functions can be used for the discriminator. The simplest solution, adopted in [14], [17] and [24], consists in using the standard binary cross-entropy (BCE) loss[1]:

$$L(\Theta, \Phi) = \mathbb{E}_{X_p}[\log(g(z_1, z_2))] + \mathbb{E}_{X_n}[\log(1 - g(z_1, z_{rnd}))], \tag{3}$$

where $\mathbb{E}_{X_p}$ and $\mathbb{E}_{X_n}$ denote the expectation over positive and negative samples, respectively. Such a metric estimates the *Jensen-Shannon* divergence between two distributions rather than the KL divergence. Consequently, this loss does not optimize the exact KL-based definition of MI, but a similar divergence between two distributions. Differently from standard MI, this metric is bounded (i.e., its maximum is zero), making the convergence of the architecture more numerically stable.

As an alternative, it is possible to directly optimize the MI with the MINE objective [15]:

$$L(\Theta, \Phi) = \mathbb{E}_{X_p}[g(z_1, z_2)] - log\Big(\mathbb{E}_{X_n}[e^{g(z_1, z_{rnd})}]\Big). \tag{4}$$

MINE explicitly computes MI by exploiting a lower-bound based on the *Donsker-Varadhan* representation of the KL divergence. The third alternative explored in this work is the Noise Contrastive Estimation (NCE) loss proposed in [16], that is defined as follows:

$$L(\Theta, \Phi) = \mathbb{E}_x\left[g(z_1, z_2) - \log\left(g(z_1, z_2) + \sum_{x_n} e^{g(z_1, z_{rnd})}\right)\right], \tag{5}$$

where the minibatch X is composed of a single positive sample and $N-1$ negative samples. In [16] it is proved that maximizing this loss maximizes a lower bound on MI.

All the aforementioned objectives are based on the idea of maximizing a discrepancy between the joint and product of marginal distributions. Nevertheless, such losses might be more or less easy to optimize within the proposed framework.

The unsupervised representations z are then used to train a speaker-id classifier in a standard supervised way. Beyond unsupervised learning, this paper explores two semi-supervised variations for learning speaker representations. The first one is

[1] The output layer must be based on a sigmoid when using BCE.

based on pre-training the encoder with the unsupervised parameters and fine-tuning it together with the speaker-id classifier. As an alternative, we jointly train encoder, discriminator, and speaker-id networks from scratch. This way, the gradient computed within the encoder not only depends on the supervised loss but also on the unsupervised objective. The latter approach turned out to be very effective, since the unsupervised gradient acts as a powerful regularizer.

Similarly to [25–28], we propose to directly process raw waveforms rather than using standard MFCC, or FBANK features. The latter hand-crafted features are originally designed from perceptual evidence and there are no guarantees that such inputs are optimal for all speech-related tasks. Standard features, in fact, smooth the speech spectrum, possibly hindering the extraction of crucial narrow-band speaker characteristics such as pitch and formants, that are important clues on the speaker identity. To better process raw audio, the encoder is based on *SincNet* [21, 22], a novel Convolutional Neural Network (CNN) that encourages the first layer to discover more meaningful filters. In contrast to standard CNNs, which learn all the elements of each filter, only low and high cutoff frequencies of band-pass sinc-based filters are directly learned from data, making SincNet suitable to process the high-dimensional audio.

3. Related Work

Similarly to this work, other attempts have recently been made to learn unsupervised representations with mutual information. In [14], a GAN that minimizes MI using positive and negative samples has been proposed for Independent Component Analysis (ICA). A similar approach can be used to maximize MI. In [16] authors proposed a method called Contrastive Predicting Coding (CPC), that learns representations by predicting the future in a latent space. It uses an autoregressive model optimized with a probabilistic contrastive loss. In [17] authors introduced DeepInfoMax (DIM), an architecture that learns representations based on both local and high-level global information.

The proposed LIM differs from the aforementioned works in the following way: DIM performs a maximization of MI between local and global representations, CPC relies on future predictions, while our method is simply based on random local sampling. Note that training using local embeddings only is very efficient since it does not require the expensive computation of a global representation as in GIM. LIM is also related with the recently-proposed methods based on triplet loss [29, 30]. Most of the previous works on triplet loss (with the exception of [31]) rely on the speaker labels [30]. Moreover, they simply maximize the Euclidean or cosine distance between speaker embeddings. LIM, instead, is based on maximizing the mutual information, thus considering a more meaningful divergence that can also capture complex non-linear relationships between the variables. Maximum Mutual Information (MMI) is often used in HMM-DNN speech recognition as a loss function [32]. This loss maximizes the MI between the acoustic probabilities and the targeted word sequence in a standard supervised framework, while LIM is used in a totally different unsupervised context that relies on local speech embeddings. Our work also uses SincNet [21, 22] (that is here used for the first time in an unsupervised framework), and extends the previous works by also addressing semi-supervised learning where encoder, discriminator, and speaker-id classifier are jointly trained from scratch. Moreover, to the best of our knowledge, this paper is the first that compares several objective functions for MI optimization in a speech task.

4. Experimental Setup

The proposed method has been evaluated using different corpora. In the following, an overview of the experimental setting is provided.

4.1. Corpora

This paper considered the TIMIT (462 spks, *train* chunk) [33], Librispeech (2484 spks), and VoxCeleb1 (1251 spks) [34] corpora. To make TIMIT and Librispeech speaker recognition tasks more challenging, we only employed 12-15 seconds of randomly selected training material for each speaker. Moreover, a set of TIMIT and Librispeech experiments have also been performed in distant-talking reverberant conditions. In this case, all the clean signals were convoluted with a different impulse response [35], that was sampled from the DIRHA dataset [36, 37]. The DIRHA corpus contains high-quality multi-room and multi-microphone impulse responses, that were measured in a domestic environment with a considerable reverberation time of $T_{60} = 0.7s$. This way, we are able to provide experimental evidence in a much more challenging acoustic scenario and we can introduce a channel effect that is not natively present in the clean TIMIT and Librispeech corpora. To study our approach using a more standard speaker recognition dataset, we also employed the VoxCeleb1 corpus (using the provided lists).

4.2. DNN Setup

The waveform of each speech sentence was split into chunks of 200 ms (with 10 ms overlap), which were fed into the SincNet encoder. The first layer of the encoder performs sinc-based convolutions, using 80 filters of length $L = 251$ samples. The architecture then employs two standard convolutional layers, both using 60 filters of length 5. Layer normalization [38] was used for both the input samples and for all convolutional layers. Next, two fully-connected leaky-ReLU layers [39] composed of 2048 and 1024 neurons (normalized with batch normalization [40, 41]) were applied. Both the discriminator and the speaker-id classifier are fed by the encoder output and consist of MLPs based on a single ReLU layer. Frame-level speaker classification was obtained from the speaker-id network by applying a softmax output layer, that provides a set of posterior probabilities over the targeted speakers. A sentence-level classification was derived by averaging the frame predictions and voting for the speaker which maximizes the average posterior. Training used the RMSprop optimizer, with a learning rate $lr = 0.001$, $\alpha = 0.95$, $\epsilon = 10^-7$, and minibatches of size 128. All the hyper-parameters of the architecture were tuned on TIMIT, then inherited for Librispeech and VoxCeleb as well.

The speaker verification system was derived from the speaker-id neural network using the *d-vector* technique. The *d-vector* [34, 42] was extracted from the last hidden layer of the speaker-id network. A speaker-dependent d-vector was computed and stored for each enrollment speaker by performing an L2 normalization and averaging all the d-vectors of the different speech chunks. The cosine distance between enrolment and test d-vectors was then calculated, and a threshold was then applied on it to reject or accept the speaker. Note that to assess our approach on a standard open-set speaker verification task, all the enrolment and test utterances were taken from a speaker pool different from that used for training the speaker-id DNN.

5. Results

This section summarizes our experimental activity on speaker identification and verification.

5.1. Speaker Identification

Tab. 1 reports the sentence-level classification error rates achieved with binary cross-entropy (BCE), MINE, Noise Constructive Estimation (NCE), and the triplet loss used in [30].

	TIMIT		Librispeech	
	CNN	SincNet	CNN	SincNet
Unsupervised-Trip. Loss	2.84	2.22	1.46	1.33
Unsupervised-MINE	2.15	1.36	1.43	0.94
Unsupervised-NCE	2.05	1.29	1.14	0.82
Unsupervised-BCE	1.98	**1.21**	1.12	**0.75**

Table 1: *Classification Error Rate (CER%) obtained on TIMIT (462 spks) and Librispeech (2484 spks) speaker-id tasks using LIM embeddings learned with various objective functions.*

The table highlights that our LIM embeddings contain information on the speaker identity, leading to a CER(%) ranging from 2.84% to 1.21% in all the considered settings. It is worth noting that mutual information losses (i.e., MINE, NCE, BCE) outperform the triplet loss. This result suggests that better embeddings can be derived with a divergence measure more meaningful than the simple cosine distance. The best performance is achieved with the standard binary cross-entropy. Similar to [17], we have observed that this bounded metric is more stable and more easy to optimize. Both MINE and NCE objective are unbounded and their value can grow indefinitely during training, eventually causing numerical issues. The performance achieved with Librispeech is better than that observed for TIMIT. Even though the former is based on more speakers, its utterances are on average longer than the TIMIT ones. The table also shows that SincNet outperforms a standard CNN. This confirms the promising achievements obtained in [21, 22] in a standard supervised setting. SincNet, in fact, converges faster and to a better solution, thanks to the compact sinc filters that make learning from high-dimensional raw samples easier.

Tab. 2 extends previous speaker-id results to other training modalities, including supervised and semi-supervised learning in both clean and reverberant acoustic conditions.

	TIMIT		Librispeech	
	Clean	Rev	Clean	Rev
Supervised	0.85	34.8	0.80	17.1
Unsupervised-BCE	1.21	28.2	0.75	15.2
Semi-supervised-pretr.	0.69	25.4	0.56	9.6
Semi-supervised-joint	**0.65**	**24.6**	**0.52**	**9.3**

Table 2: *Classification Error Rate (CER%) obtained on speaker-id with supervised, unsupervised and semi-supervised modalities in clean and reverberat conditions.*

From the table, it emerges that the results achieved when feeding the classifier with our speaker embeddings (*unsupervised-BCE*) are often better than those obtained with the standard supervised training (*supervised*). The gap becomes more evident when we pass from unsupervised to semi-supervised learning. In particular, the joint semi-supervised framework (i.e., the approach that jointly trains encoder, discriminator, and speaker classification for scratch) yields the best performance, surpassing the performance obtained when pre-training the encoder and then fine-tuning it with the supervised task (*Semi-supervised-pretr.*). The internal representations discovered in this way are influenced by both the supervised and the unsupervised loss. The latter one acts as a powerful regularizer, that allows the neural network to find robust features. The results also show a significant performance degradation in distant-talking acoustic conditions. The presence of considerable reverberation and the introduction of channel/microphone variabilities, in fact, make speaker-id particularly challenging.

5.2. Speaker Verification

We finally extend our validation to speaker verification on the VoxCeleb corpus. Table 3 compares the Equal Error Rate (EER%) achieved using our best system (*Semi-supervised-pretr.*) with some previous works on the same dataset.

	EER (%)
GMM-UBM [34]	15.0
I-vectors + PLDA [34]	8.8
CNN [34]	7.8
CNN + intra-class + triplet loss [43]	7.9
SincNet [21]	7.2
SincNet+LIM (proposed)	**5.8**

Table 3: *Equal Error Rate (EER%) obtained on speaker verification (using the VoxCeleb corpus).*

The proposed model reaches an EER(%) of 5.8% and outperforms other systems such as an I-vector baseline [34, 44], a standard CNN [34], and a CNN based on combination of intra-class and triples loss [43]. Finally, LIM outperforms a standard SincNet model trained in a fully supervised way [21]. This result confirms the effectiveness of the proposed approach even in an open-set text-independent speaker verification setting.

6. Conclusion

This paper proposed a method for learning speaker embeddings by maximizing mutual information. The experiments have shown promising performance on speaker recognition and have highlighted better results when adopting the standard binary cross-entropy loss, that turned out to be more stable and easier to optimize than other metrics. It also highlighted the importance of using SincNet, confirming its effectiveness when processing raw audio waveforms. The best results are obtained with end-to-end semi-supervised learning, where an ecosystem of neural networks composed of an encoder, a discriminator, and a speaker-id must cooperate to derive good speaker embeddings. Our achievement can be easily combined with other recent findings in speaker recognition. For instance, it is possible to use LIM to extract semi-supervised x-vectors. We can also improve it by employing an attention mechanism that weights the contribution of each time frame, or by combing our semi-supervised costs with other losses, such as the center loss.

7. Acknowledgment

We would like to thank D. Hjelm, T. Parcollet, and M. Omologo, for helpful comments. This research was enabled by support provided by Calcul Québec and Compute Canada.

8. References

[1] D. Yu and L. Deng, *Automatic Speech Recognition - A Deep Learning Approach*. Springer, 2015.

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[3] M. Ravanelli, *Deep learning for Distant Speech Recognition*. PhD Thesis, Unitn, 2017.

[4] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, “A network of deep neural networks for distant speech recognition,” in *Proc. of ICASSP*, 2017, pp. 4880–4884.

[5] M. McLaren, Y. Lei, and L. Ferrer, “Advances in deep neural network approaches to speaker recognition,” in *Proc. of ICASSP*, 2015, pp. 4814–4818.

[6] F. Richardson, D. Reynolds, and N. Dehak, “Deep neural network approaches to speaker and language recognition,” *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[8] Y. Bengio, P. L., D. Popovici, and H. Larochelle, “Greedy layer-wise training of deep networks,” in *Proc. of NIPS*, 2007, pp. 153–160.

[9] G. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets,” vol. 18, 2006, pp. 1527–1554.

[10] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *CoRR*, vol. abs/1312.6114, 2013.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. of NIPS*, 2014, pp. 2672–2680.

[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, Jan. 2016.

[13] D. Serdyuk, P. Brakel, B. Ramabhadran, S. Thomas, Y. Bengio, and K. Audhkhasi, “Invariant representations for noisy speech recognition,” *arXiv e-prints*, vol. abs/1612.01928, 2016.

[14] P. Brakel and Y. Bengio, “Learning independent features with adversarial nets for non-linear ica,” *arXiv e-prints*, vol. 1710.05050, 2017.

[15] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, “Mutual information neural estimation,” in *Proc. of ICML*, 2018, pp. 531–540.

[16] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *CoRR*, vol. abs/1807.03748, 2018.

[17] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, A. Trischler, and Y. Bengio, “Learning deep representations by mutual information estimation and maximization,” *arXiv e-prints*, vol. 1808.06670, 2018.

[18] D. Applebaum, *Probability and Information: An Integrated Approach*, 2nd ed. Cambridge University Press, 2008.

[19] J. B. Kinney and G. S. Atwal, “Equitability, mutual information, and the maximal information coefficient,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 9, pp. 3354–3359, 2014.

[20] L. Paninski, “Estimation of entropy and mutual information,” *Neural Comput.*, vol. 15, no. 6, pp. 1191–1253, Jun. 2003.

[21] M. Ravanelli and Y. Bengio, “Speaker Recognition from raw waveform with SincNet,” in *Proc. of SLT*, 2018.

[22] M. Ravanelli and Y.Bengio, “Interpretable Convolutional Filters with SincNet,” in *Proc. of NIPS@IRASL*, 2018.

[23] M. Ravanelli, T. Parcollet, and Y. Bengio, “The PyTorch-Kaldi Speech Recognition Toolkit,” in *Submitted to ICASSP*, 2019.

[24] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, “Deep graph infomax,” *CoRR*, vol. abs/1809.10341, 2018.

[25] D. Palaz, M. Magimai-Doss, and R. Collobert, “Analysis of CNN-based speech recognition system using raw speech as input,” in *Proc. of Interspeech*, 2015.

[26] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals, “Learning the speech front-end with raw waveform CLDNNs,” in *Proc. of Interspeech*, 2015.

[27] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, “Acoustic modeling with deep neural networks using raw time signal for LVCSR,” in *Proc. of Interspeech*, 2014.

[28] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, “Towards directly modeling raw speech signal for speaker verification using CNNs,” in *Proc. of ICASSP*, 2018.

[29] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering.” *CoRR*, vol. abs/1503.03832, 2015.

[30] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, “Deep speaker: an end-to-end neural speaker embedding system,” *CoRR*, vol. abs/1705.02304, 2017.

[31] A. Jati and P. G. Georgiou, “Neural predictive coding using convolutional neural networks towards unsupervised learning of speaker characteristics,” *CoRR*, vol. abs/1802.07860, 2018.

[32] L. Bahl, P. Brown, P. de Souza, and R. Mercer, “Maximum mutual information estimation of hidden markov model parameters for speech recognition,” in *Proc. of ICASSP*, vol. 11, 1986, pp. 49–52.

[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM,” 1993.

[34] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” in *Proc. of Interspech*, 2017.

[35] M. Ravanelli and M. Omologo, “Contaminated speech training methods for robust DNN-HMM distant speech recognition,” in *Proc. of Interspeech 2015*, pp. 756–760.

[36] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, “The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments,” in *Proc. of ASRU 2015*, pp. 275–282.

[37] M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo, “Impulse response estimation for robust speech recognition in a reverberant environment,” in *Proc. of EUSIPCO*, 2012, pp. 1668–1672.

[38] J. Ba, R. Kiros, and G. E. Hinton, “Layer normalization,” *CoRR*, vol. abs/1607.06450, 2016.

[39] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in *Proc. of ICML*, 2013.

[40] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. of ICML*, 2015, pp. 448–456.

[41] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, “Batch-normalized joint training for DNN-based distant speech recognition,” in *Proc. of SLT*, 2016.

[42] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in *Proc. of ICASSP*, 2014, pp. 4052–4056.

[43] N. Le and J. Odobez, “Robust and discriminative speaker embedding via intra-class distance variance regularization,” in *Proc. of Interspeech*, 2018, pp. 2257–2261.

[44] A. K. Sarkar, D. Matrouf, P. Bousquet, and J. Bonastre, “Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification,” in *Proc. of Interspeech*, 2012, pp. 2662–2665.